The Dissertation Committee for Xochitl Chamorro Morgan Certifies that this is the approved version of the following dissertation:

Eukaryotic Transcriptional Regulation: From Data Mining to Transcriptional Profiling

Committee:

Vishwanath R. Iyer, Supervisor

Orly Alter

Nigel S. Atkinson

Edward M. Marcotte

Phillip W. Tucker

Eukaryotic Transcriptional Regulation: From Data Mining to Transcriptional Profiling

by

Xochitl Chamorro Morgan, A. A.; B. S.

Dissertation

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

Doctor of Philosophy

The University of Texas at Austin

December, 2008

Dedication

For Mom, Dad, and Lisa: I could not have done this without your love and support. For Gabriel, who believed and never let me walk away. For Sheba and Maya, who kept me smiling.

Acknowledgements

I would like to thank my advisor, Dr. Vishwanath Iyer, for his scientific guidance and support throughout my graduate studies. I would also like to thank my committee members, particularly Dr. Edward Marcotte and Dr. Orly Alter, for their advice and helpful suggestions. I further thank former and current members of the Iyer and Marcotte labs for all of their help and support. In particular, I am grateful to the following people:

Dr. Patrick Killion, who gave plethora of helpful advice over the years.

Dr. Sushma Shivaswamy, who knew how to do everything and patiently answered all my questions.

Dr. Jian Gu, who taught me many techniques and helped me with statistics.

Dr. Jongwan Kim, Dr, Zhanzhi Hu, and Bum Kyu Lee, who gave advice and suggestions.

Dr. Zhihua Li, who provided yeast strains

Ryan McDaniell, who shared his PCR expertise.

Kris McGary, who helped me use the yeast interaction network.

Dr. Jonathan Davies, who showed me how to do my first microarray.

David Parr and Anna Battenhouse, who provided computer support.

Lisa Burke, who proofread.

Eukaryotic Transcriptional Regulation: From Data Mining to Transcriptional Profiling

Publication No._____

Xochitl Chamorro Morgan, Ph.D.

The University of Texas at Austin, 2008

Supervisor: Vishwanath R. Iyer

Survival of cells and organisms requires that each of thousands of genes is expressed at the correct time in development, in the correct tissue, and under the correct conditions. Transcription is the primary point of gene regulation. Genes are activated and repressed by transcription factors, which are proteins that become active through signaling, bind, sometimes cooperatively, to regulatory regions of DNA, and interact with other proteins such as chromatin remodelers.

Yeast has nearly six thousand genes, several hundred of which are transcription factors; transcription factors comprise around 2000 of the 22,000 genes in the human genome. When and how these transcription factors are activated, as well as which subsets of genes they regulate, is a current, active area of research essential to understanding the transcriptional regulatory programs of organisms. We approached this problem in two divergent ways: first, an *in silico* study of human transcription factor combinations, and second, an experimental study of the transcriptional response of yeast mutants deficient in DNA repair.

First, in order to better understand the combinatorial nature of transcription factor binding, we developed a data mining approach to assess whether transcription factors whose binding motifs were frequently proximal in the human genome were more likely to interact. We found many instances in the literature in which over-represented transcription factor pairs co-regulated the same gene, so we used co-citation to assess the utility of this method on a larger scale. We determined that over-represented pairs were more likely to be co-cited than would be expected by chance.

Because proper repair of DNA is an essential and highly-conserved process in all eukaryotes, we next used cDNA microarrays to measure differentially expressed genes in eighteen yeast deletion strains with sensitivity to the DNA cross-linking agent methyl methane sulfonate (MMS); many of these mutants were transcription factors or DNA-binding proteins. Combining this data with tools such as chromatin immunoprecipitation, gene ontology analysis, expression profile similarity, and motif analysis allowed us to propose a model for the roles of Iki3 and of YML081W, a poorly-characterized gene, in DNA repair.

Table of Contents

List of Tables

List of Figures

ABBREVIATIONS

aa-dUTP	amino allyl dUTP
cDNA	Complimentary DNA
ChIP-chip	Chromatin immunoprecipitation with microarray
Cy3	Cyanine 3
Cy5	Cyanine 5
ddH2O	distilled deionized water
DEPC H2O	Diethylpyrocarbonate-treated water
ESR	Environmental stress response
GFP	Green fluorescent protein
HR	Homologous recombination
IP	Immunoprecipitation
LAD	Longhorn microarray database
MMS	Methyl methane sulfonate
MRX	The DNA damage-sensing complex, containing Mre11, Rad50, and Xrs2
NHEJ	Non-homologous end joining
ORF	Open reading frame
PCR	Polymerase chain reaction
PWM	Position weight matrix
RT-PCR	Reverse transcription – polymerase chain reaction
SSC	Saline sodium citrate
SDS	Sodium dodecyl sulfate
TF	Transcription factor
WT	Wild type strain

Chapter 1: Introduction

EUKARYOTIC GENE REGULATION

Genes

A eukaryotic genome contains instructions for manufacturing thousands or tens of thousands of proteins. A gene, a sequence of DNA, is transcribed into a messenger RNA, which is translated into a protein; cells are made of proteins and organisms are made of cells. Yet the majority of DNA in eukaryotic genomes does not code for proteins. Much of the human genome originated from retroviral transposons [1]. Repeats of non-coding DNA called telomeres protect the ends of chromosomes, while repeats called centromeres are the sites of kinetochore and thus mitotic spindle attachment. A relatively small but very important portion of non-coding DNA is used for gene regulation.

A gene alone is an inert DNA sequence; it requires hundreds of protein-protein interactions for transcription and then translation into a functional protein. Gene expression can be regulated by chromatin condensation, DNA methylation, transcription initiation, RNA stability, translational control, and degradation, but transcription initiation is the most common point of regulation [2]. Genes cannot be transcribed without regions of DNA known as cis-regulatory regions, to which specific transcription factors will bind and thus control in which tissues, under what conditions, and in what quantities genes are expressed [2]. Cis-regulatory regions very near the transcriptional start site of a gene are known as promoters; more distal cis-regulatory regions are known as enhancers if they increase transcription and silencers if they decrease it [2].

Eukaryotic promoters for protein-coding genes contain a core promoter and a collection of transcription factor binding sites. The basal promoter is a ~100 bp stretch of DNA upon which the RNA polymerase II holoenzyme complex assembles. General transcription factors, the proteins that bind basal promoters, are ubiquitously expressed, so basal promoters alone are unable to initiate transcription under specific circumstances. Neither can they initiate transcription at high levels. Specific transcription and high-level transcription is initiated by the binding of transcription factors at locations outside the basal promoter [2].

Transcription factors are proteins that specifically influence transcription. They frequently contain DNA-binding domains, protein-protein interaction domains, ligand-binding domains, and signaling domains. DNA-binding domains determine the specificity of the sequence to which a transcription factor binds. Ligand-binding and signaling domains enable cellular localization of transcription factors and their activation when they are needed. Protein-protein interactions may include increasing or decreasing association of basal transcriptional machinery with the basal promoter, forming hetero- or homodimers with other transcription factors, physically blocking the binding sites of other transcription factors, or altering the structure of chromatin [2].

Chromatin

The genome of *S. cerevisiae* is 12.5 million base pairs, while mammalian genomes such as human and mouse are even larger: 3 billion base pairs. This DNA is organized into a structure called chromatin. The base unit of chromatin is the nucleosome, in which 146 base pairs of DNA are wrapped around a protein octamer consisting of a pair of each of four histones: H3, H4, H2A, and H2B. Each histone has a fold domain and an N-terminal tail. The fold domain keeps most of the DNA in the core of the histone octamer, where it is inaccessible to other DNA-binding proteins. The N-

terminal tails of nucleosomes facilitate coiling of chromatin into higher-order structures, which further mask DNA. Histone tails are rich in lysine and arginine, so they are positively charged to attract the negatively charged backbone of DNA [3]. These tails can be acetylated by transcriptional co-activators to promote transcription or deacetylated by co-repressors to repress transcription.

There are several ATP-dependent chromatin-remodeling enzymes. SWI/SNF family members act by perturbing the nucleosome core, while ISW1 family members cause the positions of nucleosomes to shift [3]. Histones can also be phosphorylated, methylated, and ubiquitinated [4]; in general, modification will result in chromatin structure that is more open or more closed. Closed chromatin prevents expression of all genes except those with precise positive regulation [3], while open chromatin is associated with increases in gene expression.

Transcription Factors

Transcription factors are the proteins that bind to cis-regulatory regions to enhance or repress transcription. They may expose gene promoters to make them accessible by RNA polymerase II, or they may cause bending of chromatin to bring together widely-separated regions. The RNA polymerase II holoenzyme contains a mediator structure that incorporates information received from transcriptional activators and repressors [4]. Transcription factors necessary for the expression of genes required in all cells and tissues tend to be ubiquitously expressed, while those factors that control genes needed only under specific circumstances are tightly regulated. Transcription factors may not be expressed at all until some condition, such as an embryonic developmental phase, causes their transcription. In contrast, like NFκB, they may exist in an inactive form until needed, when they are activated and translocated to the nucleus [5] .

Scope

The implications of gene regulation are far-reaching. Gene expression levels have profound effects on phenotype: it is thought that the phenotypic differences between humans and chimpanzees are due more to gene expression levels than to difference in protein sequence [6, 7]. Furthermore, the consequences of misregulation, such as uncontrolled cell proliferation, are dire. Finally, when cell disequilibrium causes human disease, knowledge of the transcriptional network contributing to the irregular state of the cells is invaluable for the development of pharmaceuticals.

To fully understand a transcriptional program, it is necessary to know the target genes, all the various transcription factors that regulate them, the circumstances under which the transcription factors bind, and how the binding affects transcript levels of the gene. DNA microarrays are a particularly useful tool for studying gene regulation, as they allow quantification of expression of all genes in a population of cells at a point in time.

Microarray Background

Photolithographic printing for use in microarrays was pioneered by Affymetrix (Santa Clara, CA) in 1991 [8], and by 1994, the first cDNA collections were being developed at Stanford. cDNA microarrays were first used to measure gene expression patterns in 1995 [9], Affymetrix released the first commercial microarrays in 1996, and the first genome-wide yeast expression studies took place in 1997 [10]. By 2000, microarrays were being used to detect gene expression signatures of cancer [11, 12]. The first whole human genome array was created in 2004 [13].

In general, microarrays consist of a chip to which a library of DNA sequences is affixed. These DNA sequences typically correspond to genes or parts of genes, but they may also be promoters or other regions of interest. Biological samples are labeled with

fluorescent dye and hybridized to the chip; nucleic acid abundance in samples correlates with intensity of fluorescence, which is measured with a laser scanner. Most microarray platforms are designed for simultaneous hybridization of two different biological samples; a notable exception is Affymetrix arrays, which are single-channel. Because microarrays allow comparison of the levels of an organism's gene expression for thousands of genes between any two samples, they are particularly useful for studying changes in gene expression over time, expression differences between tissues, or expression comparisons before and after treatment. Microarrays are also used to quantify DNA pulled down by chromatin immunoprecipitation (ChIP-chip) or to detect copy number variations by comparative genome hybridization.

There are two commonly-used types of arrays: oligonucleotide arrays and spotted arrays. Oligonucleotide arrays are created by *in situ* synthesis of the oligo on the surface of the chip by photolithography. For each transcript of interest, a set of probes is designed containing pairs of both perfect-match and mismatch oligos. Affymetrix yeast arrays contain 11 probe pairs per transcript, and the oligos are 25 base pairs long, while Nimblegen (Madison, WI) arrays range from 6-20 probes per transcript and use 60-mer oligonucleotides.

To generate spotted cDNA arrays, genes or other regions of interest are amplified by PCR, precipitated, and dried down until printing. When it is time to print, the DNA is rehydrated with water or SSC. A robot fitted with hollow-tip pins dips the pins into the 384-well plate containing the DNA, then spots the DNA onto polylysine or polyamine slides placed on the robot's platter. The pins are washed and dried between each DNA load.

One notable difference between spotted and oligonucleotide arrays is the size of transcripts of interest – the longest probes currently available on a commercial

oligonucleotide array are 60 base pairs, whereas spotted array transcripts can be hundreds of base pairs in length, making them less sensitive to hybridization changes due to gene polymorphisms [14]. Because spotted arrays do not require photolithographic masks, they are more cost-effective.

Microarray Applications

Using microarrays, we can subject cells to some perturbation and determine the genes whose expression increases, genes whose expression decreases, and genes whose expression is unaffected. We can further determine whether a transcription factor binds to a DNA sequence by cross-linking it to its chromatin, shearing the DNA, immunoprecipitating the transcription factor, and hybridizing the immunoprecipitated DNA to a microarray. Because the binding of one transcription factor may not be sufficient to cause functional changes [15-17], chromatin immunoprecipitation data provides the most information when combined with gene expression data. The ability to assess transcription factor binding and gene expression on a large scale allows powerful analysis of transcriptional programs and construction of regulatory networks.

Because genes tend to be regulated by many different transcription factors and their binding is often combinatorial, we have first developed a method to predict transcription factor cooperation in the human genome by data mining for frequently transcription factors with frequently co-occurring binding motifs. Because the DNA damage repair mechanisms of yeast and higher eukaryotes are highly conserved, we have also performed transcriptional profiling on novel MMS-sensitive yeast deletion mutants to learn more about their roles in DNA damage repair.

Chapter 2: Predicting Combinatorial Binding of Transcription Factors to Regulatory Elements in the Human Genome by Association Rule Mining

The work in this chapter was the result of a collaboration between Xochitl Morgan and Shulin Ni. It was published in *BMC Bioinformatics* [18].

ABSTRACT

Cis-acting transcriptional regulatory elements in mammalian genomes typically contain specific combinations of binding sites for various transcription factors. Although some cis-regulatory elements have been well studied, the combinations of transcription factors that regulate normal expression levels for the vast majority of the 20,000 genes in the human genome are unknown. We hypothesized that it should be possible to discover transcription factor combinations that regulate gene expression in concert by identifying over-represented combinations of sequence motifs that occur together in the genome. In order to detect combinations of transcription factor binding motifs, we developed a data mining approach based on the use of association rules, which are typically used in market basket analysis. We scored each segment of the genome for the presence or absence of each of 83 transcription factor binding motifs, then used association rule mining algorithms to mine this dataset, thus identifying frequently occurring pairs of distinct motifs within a segment. Support for most pairs of transcription factor binding motifs was highly correlated across different chromosomes although pair significance varied. Known true positive motif pairs showed higher association rule support, confidence, and significance than background. Our subsets of high-confidence, high-significance mined pairs of transcription factors showed enrichment for co-citation in PubMed abstracts relative to all pairs, and the predicted associations were often readily verifiable in the

literature. In conclusion, functional elements in the genome where transcription factors bind to regulate expression in a combinatorial manner are more likely to be predicted by identifying statistically and biologically significant combinations of transcription factor binding motifs than by simply scanning the genome for the occurrence of binding sites for a single transcription factor.

BACKGROUND

Substantial differences of phenotype can be primarily the result of differences in gene expression levels rather than in protein structure. Genes are dynamically regulated, primarily at the transcriptional level, by protein transcription factors that bind DNA at cis-regulatory regions to activate or repress expression. Mammalian cis-regulatory regions range in length from the 60 bp human muSK enhancer [19] to the 450 bp human TGFβ enhancer [20] to the 1100 bp enhancer of murine Pax6 [21], but they are generally a few hundred base pairs in length. Enhancers contain binding sites for transcription factors, sometimes for a single factor and sometimes for many [22]. A detailed understanding of the transcriptional regulatory programs of any organism requires knowledge of the binding sites of transcription factors, the circumstances and cellular conditions under which these transcription factors bind to their targets, and the genes that are regulated by combinations of transcription factors.

Cis-regulatory regions for most of the approximately 20,000 protein-coding genes encoded in the human genome have not yet been characterized [23]. Transcription factor binding sites, and thus cis-regulatory regions, can be identified using high-throughput methods such as ChIP-chip [24-27], but there are more than 2000 transcription factors encoded in the human genome [28, 29]. This diversity of transcription factors, coupled with the fact that many are likely to be expressed and to combinatorially regulate target genes in a developmental, cell-, or tissue-specific manner, makes experimental

identification of cis-regulatory regions challenging even with genome-wide ChIP-chip. Computational identification of cis-regulatory motifs based on signatures of their presence in the genomic sequence is an attractive alternative.

A major class of computational methods for identifying regulatory elements relies on the occurrence of TF binding sites in close proximity within regulatory elements. For example, the stripe 2 enhancer of the *even-skipped (eve)* gene in *Drosophila melanogaster* has twenty binding sites for four TFs within an area of roughly 600 bp [30]. The *knirps* gene of *Drosophila* is regulated by two enhancers containing six binding sites each for the transcription factors *bicoid* and *caudal* as well as two *hunchback* sites [31]. The HS2 enhancer of the human β-globin locus contains four NF-E1 binding sites and 2 CACC boxes within 250 bp [32], while a 300 bp region near the interleukin 2 transcriptional start site contains multiple binding sites for Ap-1 and Oct1 as well as sites for NFκB and NFAT [33]. Thus, the density of TF binding sites may be used as a means to locate cis-regulatory regions computationally [34].

Computational location of cis-regulatory modules by clustering of transcription factor binding sites has been implemented in genomes ranging from yeast [35] to human [36]. Previous approaches include "sliding window" [37-39] to Hidden Markov models [34, 36] to position weight matrix clustering [40-44], while clusters have been defined both homotypically [37, 39] and heterotypically [38, 45-47]. These computational methods have been used to locate many cis-regulatory regions and novel target genes, notably in *Drosophila*. One limitation of these heterotypic clustering methods is the need to know which combinations of transcription factors should define the heterotypic clusters.

Numerous transcription factors are known to cooperate in certain contexts; for example, it is known that many genes involved in inflammation are regulated by Ap-1

and NFκB [48]. Similarly, interactions between PU.1 and GATA family TFs mediate cell differentiation in B-cell development [49]. Prediction of transcription factor cooperativity has been carried out in yeast [50, 51] and human [52] genomes, but elucidation of the entire network of transcription factors that cooperate with one another in cis-regulatory regions is far from complete. In order to better define biologically relevant, heterogeneous combinations of transcription factors, we have developed an association rule data mining approach to search genome sequence information and identify over-represented adjacent motifs for transcription factor binding. Predicting transcription factor cooperation by data mining using association rules has previously been attempted in yeast as well as *C. elegans* and human chromosome 22 [53-55], but these attempts have been limited to mining known promoters [53] or repetitive elements such as microsatellites [54, 55] rather than applied to the entire human genome.

Association rule data mining [56] was originally used in market basket analysis to determine which items are frequently purchased together. Basket analysis uses a database of transactions in which each tuple is a list of items purchased in one customer's transaction. Mining seeks to discover rules such as "spaghetti $\Rightarrow$ parmesan cheese," meaning "People who buy spaghetti also often buy parmesan cheese." Association rules can be formally described as follows: [56]

- $I = \{i_1, i_2 \ldots i_n\}$ is a set of literals called items.
- D is a set of transactions. Each transaction T is a set of items such that $T \subseteq I$.
- A transaction T contains X, a set of items in I, if $X \subseteq T$.
- An association rule is an implication of $X \Rightarrow Y$, where $X \subset I$, $Y \subset I$, and $X \cap Y = \varnothing$.
- C is the confidence of a rule $X \Rightarrow Y$ in transaction set D if $c\%$ of transactions in D that contain X also contain Y. It is also known as the conditional probability of Y

given X, or P(Y|X).

- *S* is the support of rule $X \Rightarrow Y$ in set *D* if *s*% of transactions in *D* contain both X and Y. It is also known as the joint probability of both X and Y, or $P(X \cap Y)$.

If a rule $X \Rightarrow Y$ has high confidence, it is likely that transactions containing *X* will likely also contain *Y*. However, the existence of such a rule does not by itself imply any causal relationship between *X* and *Y*.

Determining over-represented transcription factor partners may help to reveal biological roles for less well-studied transcription factors. Therefore, in our studies, we used data mining to determine whether two transcription factors whose experimentally determined binding motifs were frequently proximal to one another were also likely to have biologically meaningful interactions. For example, the rule "Nuclear Factor Kappa $B \Rightarrow$ Ap-1" would indicate "Where there is a motif for NFκB, there is often also an Ap-1 motif." To allow application of association rules to transcription factor motifs in the human genome, we divided the genome into segments and scored each segment for the presence or absence of each of 83 transcription factor binding motifs (Figure 2.1). Thus, the set of 83 motifs becomes *I*, each individual transcription factor binding motif becomes an item, and each small segment of genome becomes a transaction *T* whose contents *X* are the motifs located within.

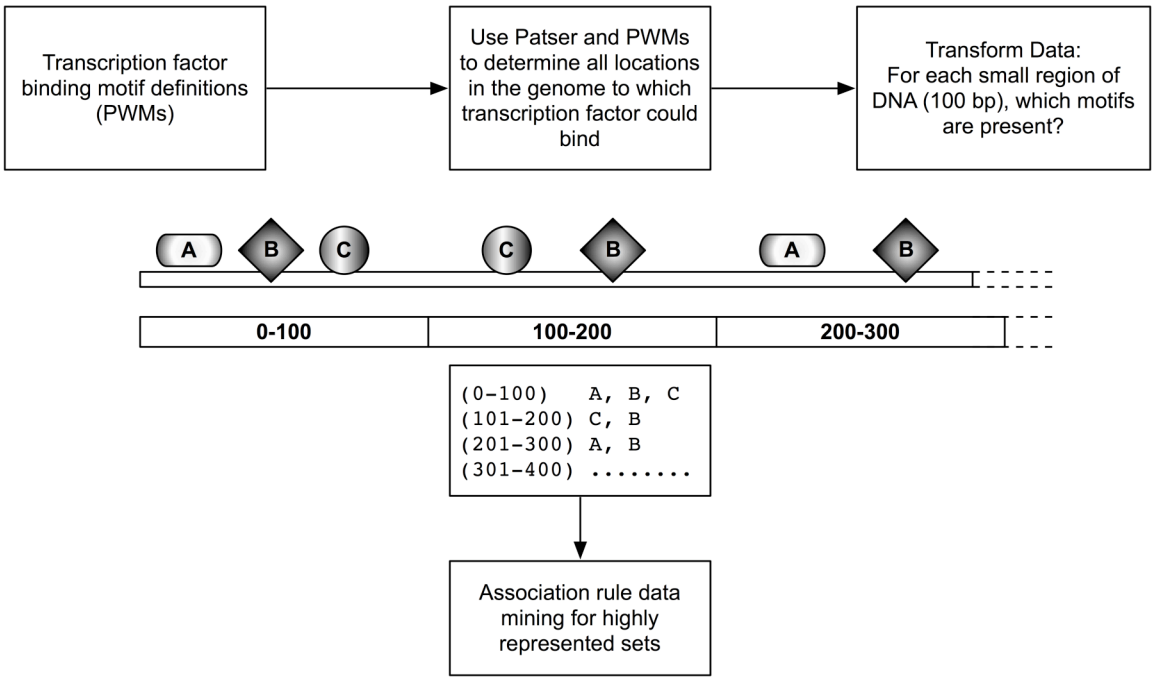


Figure 2.1: Overview

Patser is used to map all possible binding sites in the genome for each of 83 position weight matrices (PWMs) from Transfac. The genome is divided then scored 100 bp at a time for the presence or absence of each PWM, and association rules are used to mine the genome for frequently co-occurring pairs.

Results

Our major aim was to determine whether a pair of transcription factors whose motifs were frequently near one another were more likely to have a biological association than a pair of transcription factors whose motifs were not. In order to test this hypothesis, we located all possible binding sites in the human genome for the position weight matrices (PWMs) of each of 83 transcription factors (Supplemental Table 2.1). We then divided the genome into 100 bp regions and used association rule data mining to calculate support and confidence for each transcription factor pair in the human genome.

Straightforward association rule mining that simultaneously considers all motif positions discovers high numbers of transcription factor pairs that bind identical or highly similar motifs. For example, two different transcription factors A and B may both bind to the motif "CACGTG", so the confidence *C* of the rule $A \Rightarrow B$ will be 100%. Similarly, if A binds to "CACGTG" and B binds to "CACGTGA," this high overlap between binding motifs will result in the confidence being very high while the rule is neither interesting nor surprising, although it may still be biologically valid. To avoid discovery of enriched overlapping motifs, for each transcription pair AB, all overlapping binding sites between A and B were removed before calculating support and confidence (Figure 2.2). We also calculated a *P*-value based on the hypergeometric probability of observing the association between A and B by chance.

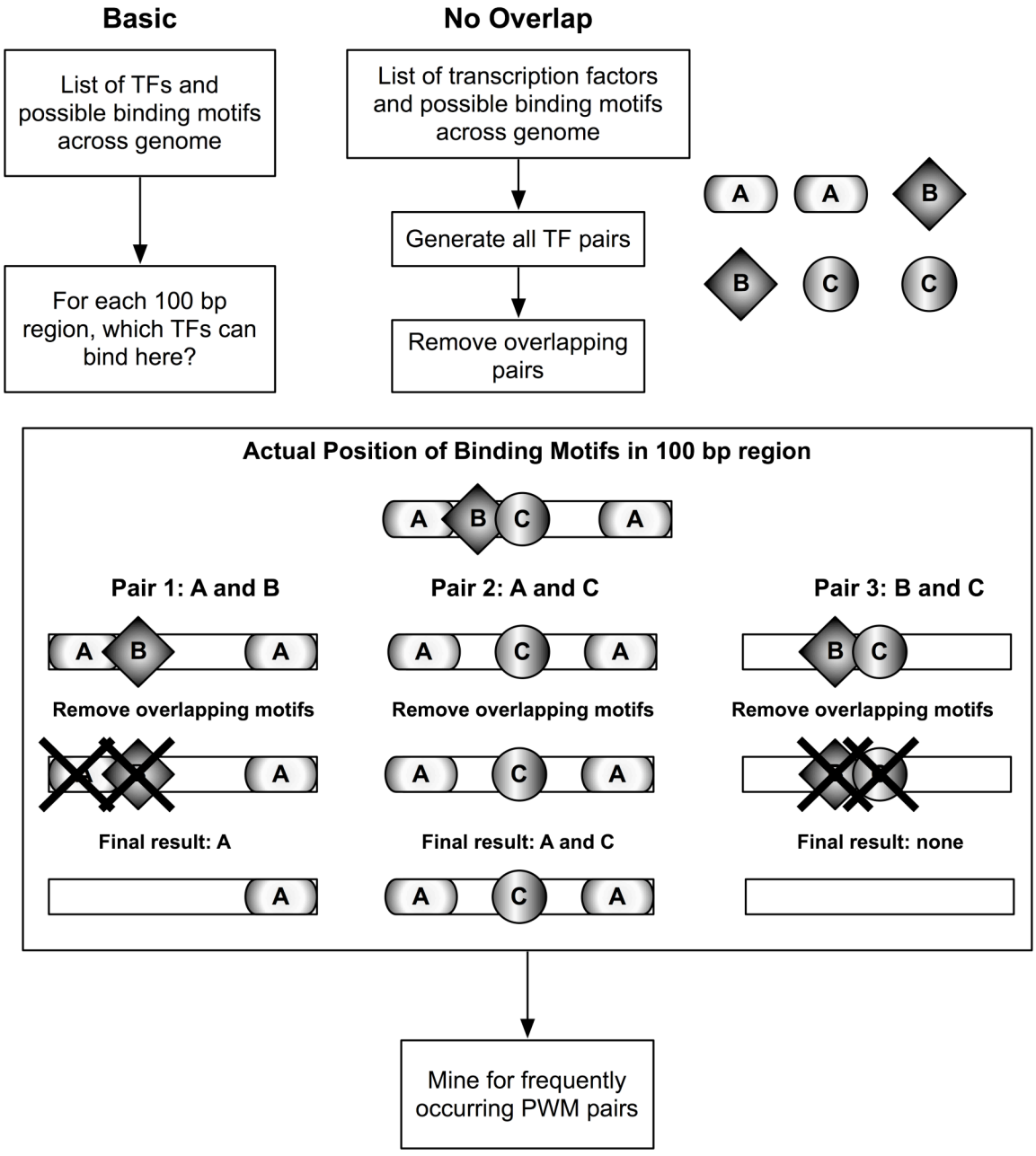


Figure 2.2: Mining without overlap

In order to avoid enriching primarily for TF pairs that bind similar motifs, the genome is mined once for each pair AB. All overlapping motifs between A and B are removed before calculating support, confidence, and *P*-values, then restored upon subsequent iterations.

In order to determine whether biologically significant associations between PWMs arise in promoter regions, we applied the same pairwise mining algorithm to the subset of the genome that was 1 kb upstream of the transcriptional start site of all human RefSeq genes [57]. Because transcription factor function is often phylogenetically conserved, we also examined whether the combinations we identified by mining the human genome were identifiable in the mouse genome; we performed identical pairwise mining for significant associations among the same 83 transcription factors on mouse chromosome 1.

Identifying Meaningful TF Pairs

Due to the size of the human genome and the tendency of PWMs to match at a large number of genomic locations, all TF pairs showed some co-occurrence. This support for possible transcription factor PWM pairs ranged from 9×10^{-6} to 0.2. Support for the association of a given pair of transcription factors was highly conserved, not only between promoters and the entire genome (Figure 2.3A), but also between the human chromosomes and mouse chromosome 1 (Figure 2.3C) and between individual human chromosomes (Figure 2.3B, Figure 2.3D), suggesting that the associations revealed by mining are biologically relevant.

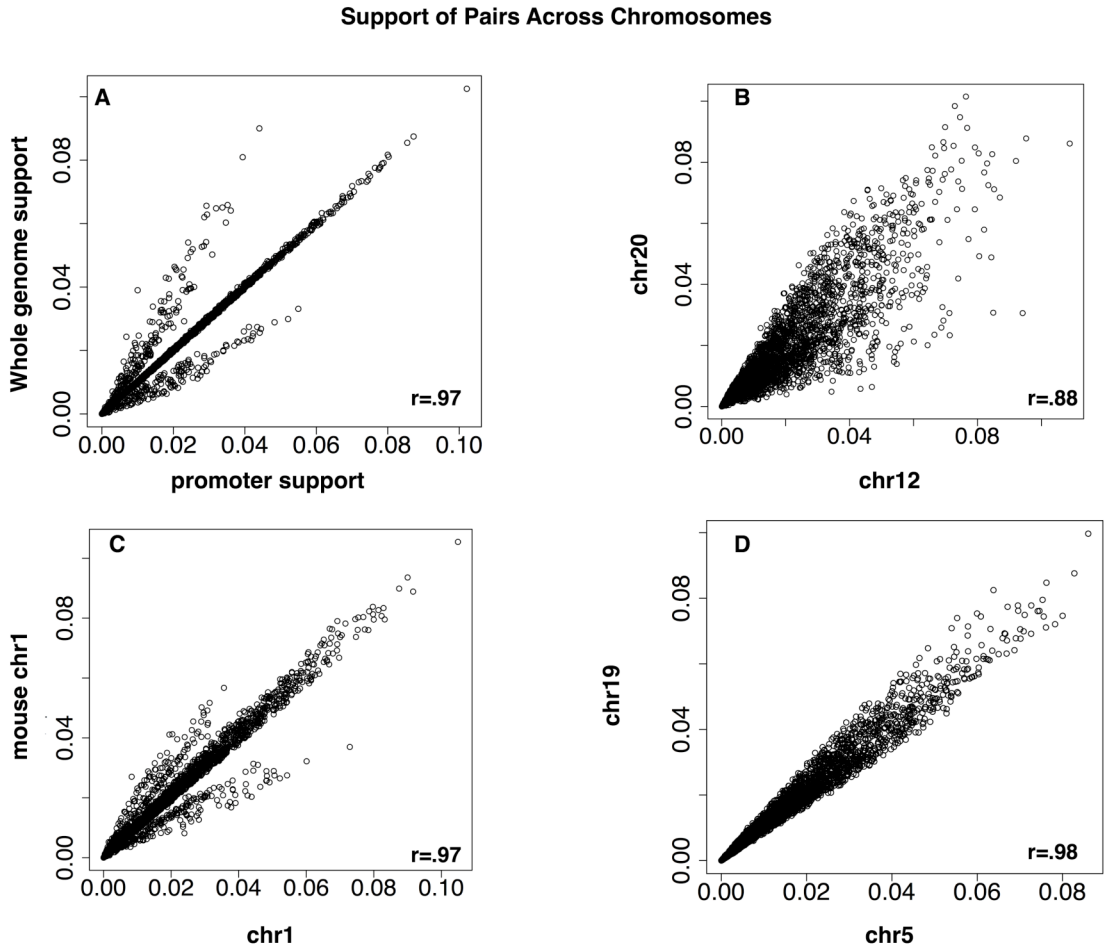


Figure 2.3: Support of TF pairs across chromosomes

The support of a given TF pair is highly correlated between chromosomes (B, D). This is also true for support in promoter regions versus the entire human genome (A) as well as support between human and mouse chromosomes (C).

Association rules with the highest support and confidence are typically regarded as being interesting; however, if two different transcription factors each have large numbers of independent binding motifs in the genome, they could appear to be associated with high support values merely by chance. To minimize this possibility and to select those TF pairs occurring more frequently than by random chance and thus likely to be biologically meaningful, we additionally calculated the statistical significance (*P*-value) of observing each TF pair using the hypergeometric probability distribution. We defined the dataset "all" as the complete set of 3403 PWM pairs, and we selected three subsets with high confidence and significance for further analysis: "genomewide," "mouse," and "promoter" (Supplemental Table 2.2).

The subsets "genomewide," "promoter," and "mouse" were defined as $P < 0.05$, greater than median difference between confidence $A \Rightarrow B$ and confidence $B \Rightarrow A$. For the subset "genomewide" this was measured on the entire human genome and resulted in 66 TF pairs. For the subset "mouse," this was measured on mouse chromosome #1 and resulted in 184 pairs. For the subset "promoter," this was measured only across regions 1 kb upstream of the transcriptional start site of each RefSeq gene and resulted in 28 pairs.

The subsets of PWM pairs chosen for further inspection were of exceptionally high support and statistical significance. They co-occurred within the same short segment of DNA throughout the human genome much more often than the others, and much more frequently than expected by chance given their individual distributions. Transcription factors binding to the motifs represented by these PWMs were therefore expected to bind and jointly regulate the expression of target genes.

Microarray Verification

We hypothesized that high-support, high-significance TF pairs or their target genes might be co-expressed in microarray data more often than other pairs. Therefore, we calculated the Pearson correlations of expression for all genes across 4742 human microarrays from the Stanford Microarray Database, but we saw no difference between the expression correlations of selected TF pairs and all TF pairs and no difference between genes containing both members of a high-support, high-significance motif pair 1 kb upstream of the transcriptional start site and genes without (data not shown).

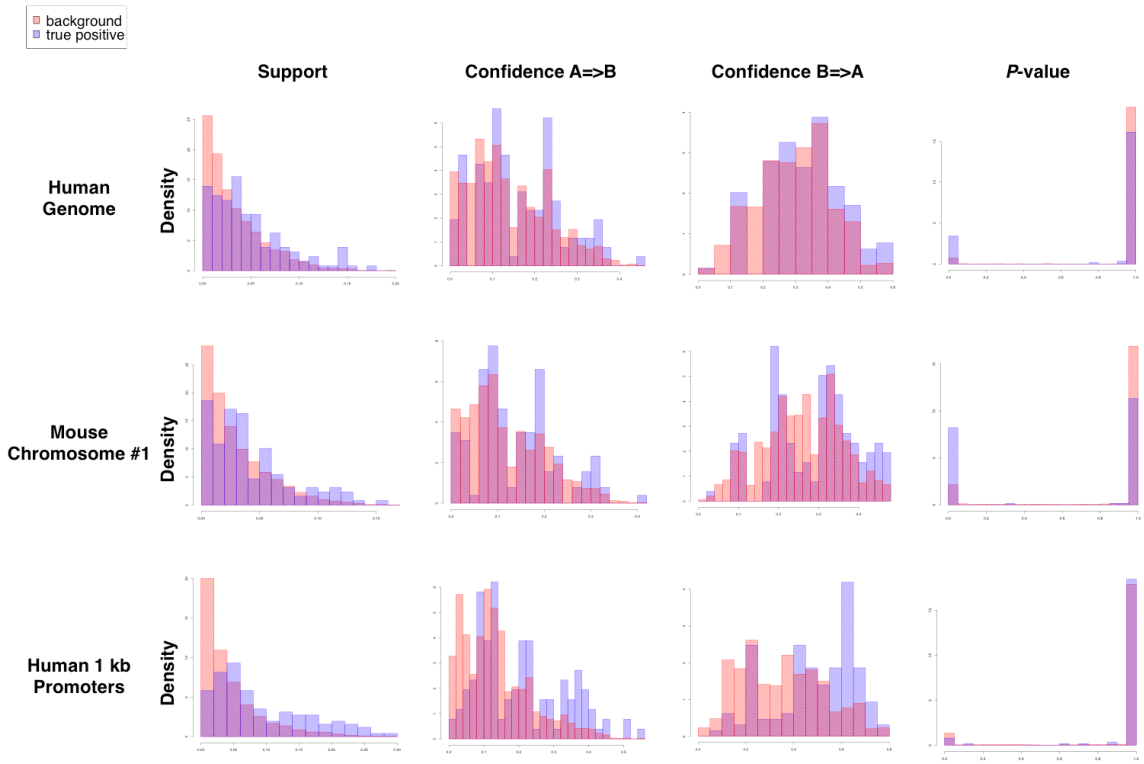


Figure 2.4: True positives vs. all pairs

Distribution histograms of support, confidence, and *P*-value for 131 true positives versus all pairs show higher support and confidence and lower *P*-values for true positives in the entire human genome, human promoter regions, and mouse chromosome #1.

Table 2.1: High-confidence TF pairs verified in the literature

Examples of high-confidence TF pairs that could be verified in the literature as co-regulators of mammalian genes.

TF pair	Gene Regulated	Source
Ap-2, p300	Mouse CITED4	[58]
Sp1, Gata2	Human PDGFβ receptor	[59]
Sp1, p300	Human ERK1	[60]
Ap-2, Egr1	Human tumor necrosis factor α, rat chromogranin B, human PNMT	[61-64]
Ap-2, NFκB	Human tumor necrosis factor α	[61]
Egr1, Elk1	Human tumor necrosis factor α	[61, 65]
Egr1, Nf1	Human tissue factor pathway inhibitor 2	[66]
Egr1, p300	Human tumor necrosis factor α	[62]
Egr1, Sp1	Human TFPI-2, human SOD, human cd95, human TNFα	[65-68]
Sp1, p53	Human Icam1	[69, 70]
Mzf1, Sp1	Human N-cadherin	[71]
Sp1, Srebp	Porcine LDL receptor, rat FAS	[72, 73]
Usf, Sp1	Rat FAS, human Top3, human liver fructose1,6 biphosphatase	[73-75]
Aml1, NFκB	Human GM-CSF	[76]
Aml1, Srebp	Human fatty acid synthase	[77]
Elk1, p300	Human tumor necrosis factor α	[65]
Gata2, NFκB	Human erythropoietin	[78]
Gata2, Sp1	Human PDGF receptor	[59]
Nf1, NFκB	Human tissue factor pathway inhibitor 2	[66]
NFκB, p300	Human I-gamma 1, mouse tapasin	[79, 80]
Pax5, p300	Human immunoglobin κ	[81]
Ap-1, NFκB	Human interleukin 6, human RANTES, human TNFα, human GM-CSF	[61, 82-84]

Verification in the Literature

We next manually examined the literature for evidence of biological associations and joint regulation of target genes by the "genomewide" and "mouse" subsets of PWM pairs that were identified by data mining. We found that many of these TF pairs were readily verifiable in the literature as true co-regulators of human and mouse genes (Table 2.1). For example the subsets "mouse" and "genomewide" both included the pair "Ap-2, Egr1." Genes known to be regulated by these two transcription factors include tumor necrosis factor α [61, 66], human phenylethanolamine N-methyltransferase [64], and rat chromogranin B [63]. The subsets "mouse" and "genomewide" contain the pair "Sp1, p53"; each has been shown to regulate ICAM-1 [69, 70]. A comparison of distributions for all pairs compared to 131 true positives collected from the literature revealed that true positive pairs exhibited higher support and confidence and lower *P*-values than did all pairs (Figure 2.4), regardless of whether the entire human genome, human promoters, or mouse chromosome 1 were mined. As an exhaustive manual analysis of the literature for all TF pairs was not feasible, we used high-throughput co-citation analysis to further assess the biological relevance of the high-support, high-confidence TF pairs.

High-Throughput Co-citation

In order to determine whether the members of a TF pair were co-cited in the literature more often than expected by chance and more often than the pairs that were not significant, we used the CoCiteStats package in R [85] to calculate PubMed co-citation rates for all TF pairs and subsets. For each pair of PWMs, CoCiteStats calculates co-citation by determining the concordance, Jaccard index, and Hubert's Γ, as well as the *P*-values for these indices, which are significant at $P < 0.05$ [86]. Concordance is a

straightforward measure of how many papers in PubMed co-cite both genes. The Jaccard index is the ratio of the number of papers containing both genes to the number of papers containing at least one of the two genes. Hubert's Γ measures the degree of association between two binary variables, ranges from -1 to 1, and can be interpreted similarly to the Pearson correlation [86]. Because papers that cite a large number of genes are less likely to contain meaningful information about interactions between any two genes cited in that paper than papers citing fewer genes, CoCiteStats also weights data for paper size (number of genes cited in a paper), gene size (number of papers that cite a gene), and both gene and paper size [86].

Figure 2.5 shows the fraction of total TF pairs with significant co-citation *P*-values ($P < 0.05$) in each dataset. Asterisks indicate a significant difference between all TF pairs and the selected subset as measured by a Chi square test. All sets indicated by "§" were significant after Bonferroni correction for multiple hypothesis testing. All three subsets showed substantially higher proportions of TF pairs enriched for low co-citation *P*-values in all cases than the set of all pairs, indicating that transcription factors binding to the PWMs that showed substantial association with one another on the genome were more likely to be co-cited in the literature, reflecting a likely biological association between them. This enrichment of "genomewide" was significant for most values at all adjustments. The subset "mouse" was enriched for significant concordances and Jaccard values when unadjusted or adjusted by paper size and was significant for all values when adjusted by both gene and paper size. The subset "promoter" was more significant after adjustments for gene size or both gene and paper size.

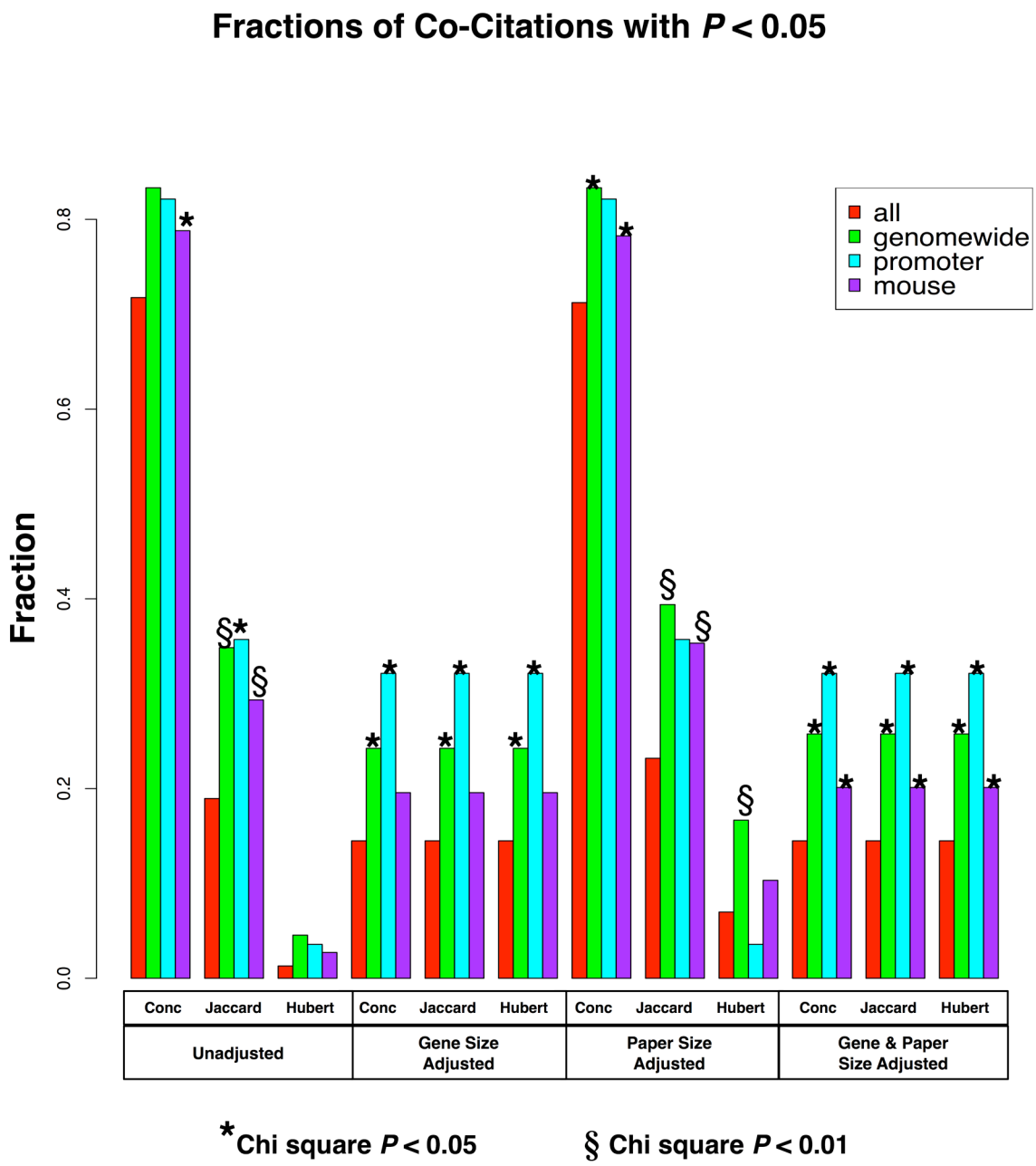


Figure 2.5: Fractions of TF pairs with significant co-citation *P*-values

Fractions of TF pairs with significant co-citation *P*-values ($P < 0.05$) in each dataset. Asterisks indicate a significant difference between all pairs and the selected subset as measured by a Chi square test. *P*-values significant after the Bonferroni correction for multiple hypothesis testing are indicated by "§".

DISCUSSION

Data mining using association rules discovered biologically meaningful cooperating TF pairs. Known true positive TF pairs showed higher support, confidence, and significance than did all pairs. Mined pairs with high significance as measured by the hypergeometric probability distribution and a large difference between confidence A=>B and confidence B=>A were frequently verified in the literature and showed enrichment of low co-citation *P*-values. We found that data mining the entire human genome was a better indicator of biological significance than was mining mouse chromosome #1, as measured by co-citation.

Given that phylogenetically conserved transcription factor binding motifs are thought to be biologically useful [87], it is interesting that 90% of the TF pairs in the subset "genomewide" were also present in the subset "mouse." Comparison of TF pairs for multiple mouse chromosomes or across more than two mammals may lead to even better results. The smaller overlap between "promoter" and "mouse" (46%) and "promoter" and "genomewide" (36%) may be due in part to differences in sequence size and nucleotide frequency; the sequence mined for "promoter" was a tenth the size of the sequence mined for "mouse" and ~1/200 the size of the sequence for "genomewide." Furthermore, the "promoter" sequence has a much higher GC content of 53% GC, while the human genome and mouse chromosome #1 are 41% GC; the PWMs used for mining have an average GC content of 46%.

Approximately 2900 of the TF pairs in our analysis were non-significant on mouse chromosome #1, human 1 kb promoter regions, or the human genome. The most confident of the remaining ~550 TF pairs may merit further study. Our estimated error rates for PWM matches located by Patser ranged from 3.5% to 61.5% with an average of

20% and a median of 18% (Supplementary Table 2.3). Pairs containing a TF with a very high error rate are less likely to be of predictive value, but most TF pairs with high confidence and significance did not have very high Patser error rates.

Our approach is novel, low-cost, and straightforward to implement. The main advantage of this approach is that the signal for the association of transcription factors is detectable using only the genome sequence and is not limited by lack of prior knowledge about physiological conditions or cell types in which the transcription factor combination may be active. Unlike clustering algorithms, which require items to be assigned to only one cluster, association rules allow items to be members of many groups and may discover these relationships. This algorithm also enables us to analyze a great number of motifs and large amount of sequence data for which Gibbs sampling is not currently feasible. One limitation of our current implementation is that we have applied it to identify only combinations of two distinct transcription factors. Although it is possible to discover associations of multiple transcription factors in the genome sequence through association rule mining, this is more computationally demanding.

As with any computational prediction, the significant challenge is verification of the predicted TF pairs. Co-citation analysis was particularly useful given that expected measures of biological association between the members of predicted TF pairs, such as correlated expression of target genes and network connectivity, were not useful. There are several possible explanations for why we did not observe correlations for significant mined TF pairs in microarray data. First, the activity of transcription factors may not be primarily regulated transcriptionally. Rather, transcription factors may require degradation of chaperones to become active, as does NFκB, or ligand binding may be needed to cause an active receptor to relocalize to the nucleus, as in the case of the estrogen receptor. While some transcription factors, such as targets of immediate early

genes, may have similar mechanisms of transcriptional activation, it is likely that many, if not most cooperating transcription factors will have diverse means of transcriptional regulation and will thus not be co-expressed. Furthermore, due to noise and the fact that transcription factors may be inactive in many cell types and experimental conditions, any co-expression signature may be lost in large amounts of microarray data even for transcription factors known to be co-expressed. For example, across the 4247 microarrays we analyzed, the Pearson correlations for Fos with JunB and Jun were -0.11 and 0.146, respectively; the correlation was -0.116 for Gata2 and Gata3 and 0.24 for Sox5 and Sox6. Thus, even for known pairs of transcription factors, there is little detectable coexpression across a large microarray dataset.

We found that genes containing significant pairs of PWMs in their promoters were no more likely to be co-regulated than a background set. One possible explanation is that our list of 4742 microarrays represented a wide variety of experimental conditions, but many of the transcription factors we studied are active only under specific conditions satisfied in only a small number of experiments. Furthermore, the short, degenerate nature of position weight matrices means that thousands of 1 kb upstream regions are likely to contain any given PWM pair. We found that each PWM was present in the upstream regions of 10,948 genes on average, while the promoter region of each gene contained an average of 70 PWMs (data not shown). Thus, any comparisons of subsets became comparisons of most genes versus most genes, making it difficult to detect a change in the distribution of correlation coefficients. Observing correlated expression of the target genes of highly supported TF pairs would be much more likely if target genes could be more rigidly defined and a subset of microarray experiments was chosen to reflect likely conditions for transcription factor activity, but choosing these experiments is nontrivial, particularly for transcription factors that have not been well-studied.

Co-citation is not without drawbacks. The fact that two proteins are cited in a paper does not necessarily mean that they interact with one another. Furthermore, well-studied proteins are likely to be overrepresented while less-studied proteins will be missed. Validation by co-expression, however, requires knowledge of target genes and conditions for transcription factor activity; this may not be known or be feasible for experimental analysis. Future experimental validation of predicted associations could be accomplished by identifying binding targets for these transcription factors by genome-wide chromatin immunoprecipitation analyses and determining joint occupancy of target promoters by predicted combinations of transcription factors. Current maps of human protein-protein interactions [88-92] may not yet define many interactions for human transcription factors or may contain high rates of false positives [93], but they are constantly improving. We anticipate that better human protein-protein interaction maps will eventually provide a superior means of assessing performance of TF pair data mining, allowing this method to be refined to reveal both novel transcription factor interactions and biological context for previously uncharacterized transcription factors.

Conclusion

Here we have described a novel genomic method for predicting biologically relevant, heterogeneous combinations of cooperating transcription factors by data mining using association rules to search genome information and identify over-represented proximal motifs. Using this approach, we show that that true positive cooperating TF pairs tend to have higher support, confidence, and significance, and that mined TF pairs with high confidence and significance are frequently verified in the literature and enriched for low co-citation *P*-values. Data mining the entire human genome enabled better discovery of biologically meaningful pairs than mining mouse chromosome #1, as measured by co-citation.

METHODS

Data Transformation

We collected 163 human position weight matrices (PWMs) from Transfac [94] and removed those which were redundant or could not be mapped to RefSeq genes [57], leaving 83 PWMs for analysis (Supplementary Table 2.1). We used Patser [95] to map all locations in the human genome assembly hg17 and in the repeat-masked human genome assembly hg18 [96] to which each transcription factor could bind with $P < 0.001$. We then divided the genome into 100 bp regions and scored each region for the presence or absence of each PWM. We chose a region size of 100 bp because it is compatible with the size of known cis-regulatory regions and large enough to contain multiple non-overlapping transcription factor binding motifs. PWMs tend towards large numbers of possible binding sites in the genome; 100 bp regions are small enough to prevent most regions from containing most motifs. We mined this matrix of genomic regions and motifs they contained for frequent itemsets, using association rules to search for $X \Rightarrow Y$ with high support S. Support and confidence were highly correlated between hg17 without repeat masking and hg18 with repeat masking (Supplementary Figure 2.1). High-support, high-confidence, significant PWM pairs were comparable between region sizes ranging from 75 bp to 225 bp, although larger region sizes yielded greater numbers of significant pairs.

Estimating Patser Error Rate for PWMs

We estimated Patser error rates for each position weight matrix by calculating its average P-value across the genome as given by Patser, multiplying this by the size of the genome minus the length of masked repeats and then dividing by total number of matches to approximate the number of overestimated Patser matches

Mining Without Overlap

In order to avoid enrichment of PWMs with highly similar binding motifs, we mined the human genome without allowing motif overlap, one motif pair at a time. That is, for each TF pair AB (83 transcription factors taken two at a time, or 3403 pairs), after all possible binding motifs for A and B respectively were identified, any overlapping A and B motifs were removed before assigning the remaining non-overlapping sites to their respective 100 bp regions (Figure 2.2). The full set of matches for each factor was restored at the beginning of each iteration, so overlaps between A and C were unaffected by overlaps between A and B. For example, if transcription factor A had a binding motif of width 5 which was present at positions 100, 130, and 150, while factor B had a binding motif of width 7 present at 102, 160, and 175, the binding sites 100A and 102B would be removed from calculations due to overlap; the remaining binding sites would still allow the region from 100 to 200 to be scored as containing A and B. After scoring each 100 bp region, we calculated association rule support (proportions of regions) for A, B, and AB for each pair on each chromosome, correcting for the proportion of the genome that was repeat-masked. Additionally, we calculated confidence for A⇒B and B⇒A and a *P*-value based on the hypergeometric probability of observing the association between A and B by chance, given the individual distributions of their binding motifs in the genome, again correcting for repeat masking. To allow phylogenetic comparison and comparison of promoters versus the entire genome, we performed identical pairwise mining on mouse chromosome 1 and on the subset of the human genome that was 1 kb upstream from the transcriptional start site of all human RefSeq genes.

Microarray Data

To determine whether transcription factor pairs with high support and high confidence were highly co-expressed, we downloaded and analyzed a dataset consisting

of 4742 human microarrays from the Stanford Microarray Database [97] and calculated the Pearson correlation for each gene pair with 100 or more experimental data points. We defined a list of potential target genes for TF pairs by scanning 1 kb upstream from the transcriptional start site of each RefSeq gene for each PWM.

True Positives

From the Compel database [98] and the literature, we collected 131 transcription factor pairs known to co-regulate mammalian genes [58-62, 66-80, 82-84, 99-151].

Software Availability

Software is available for download at http://sourceforge.net/projects/miner/.

AUTHORS' CONTRIBUTIONS

XM carried out data analysis. XM and SN wrote the computer code. XM and VI wrote the manuscript. DM provided the initial impetus for the design of this project. All authors participated in the design of the study.

ACKNOWLEDGEMENTS

We thank Orly Alter, Edward Marcotte, Patrick Killion, and Iyer lab members for advice and suggestions. This work was supported in part by a NIAAA Alcohol Training Grant and an ITR grant from the National Science Foundation.

Supplemental Table 2.1: 83 transcription factors from TRANSFAC

A list of the 83 transcription factor position weight matrices from TRANSFAC used for this analysis.

AML1_01
AP1_Q2
AP2_Q6
AP4_Q6
AREB6_03
ARNT_01
ARP1_01
ATF_01
BRN2_01
CART1_01
CDP_02
CEBP_C
CHOP_01
COUP_01
CREB_01
CREBP1_01
CREL_01
E2F_01
E2F_02
E47_02
EGR1_01
ELF1_01
ELK1_02
ER_Q6
FOXD3_01
FOXJ2_02
FREAC2_01
FREAC3_01
FREAC4_01
FREAC7_01
GATA2_01
GATA3_01
GATA_C
GRE_C
HFH3_01
HLF_01
HNF1_01
HNF4_01
HSF1_01
HSF2_01
IRF1_01
ISRE_01
LMO2COM_02
MAX_01
MEF2_03
MEF2_04
MEIS1_01
MIF1_01
MYB_Q6
MYCMAX_01
MYOD_Q6
MZF1_01
NF1_Q6
NFAT_Q6
NFE2_01
NFKAPPAB_01
NFY_01
NKX61_01
NRSF_01
OCT_C
P300_01
P53_01
PAX2_01
PAX5_01
PAX6_01
PBX1_02
RFX1_01
RORA1_01
RORA2_01
RREB1_01
SOX9_B1
SP1_Q6
SREBP1_02
SRF_Q6
SRY_02
STAT_01
TATA_01
TCF11_01
TGIF_01
TST1_01
USF_01
XBP1_01
YY1_02

Supplemental Table 2.2: The subsets "genomewide," "promoter," and "mouse"

The subsets "genomewide", "mouse", and "promoter" are defined as top 50% difference between confidence A=>B and confidence B=>A and $P < 0.05$ as measured by the hypergeometric distribution. Pairs indicated in bold have been verified in the literature.

Genomewide		
AML1_01.EGR1_01	ARNT_01.PAX2_01	**MAX_01.SP1_Q6**
AML1_01.MYCMAX_01	ARNT_01.PAX5_01	MYCMAX_01.MZF1_01
AP2_Q6.ARNT_01	ARNT_01.RREB1_01	MYCMAX_01.PAX5_01
AP2_Q6.ATF_01	ARNT_01.SP1_Q6	**MYCMAX_01.SP1_Q6**
AP2_Q6.E47_02	ARP1_01.EGR1_01	MYOD_Q6.NFKAPPAB_01
AP2_Q6.EGR1_01	ATF_01.SP1_Q6	MYOD_Q6.NRSF_01
AP2_Q6.NRSF_01	E47_02.MZF1_01	MYOD_Q6.SREBP1_02
AP2_Q6.MAX_01	**EGR1_01.ELK1_02**	MZF1_01.NRSF_01
AP2_Q6.MYCMAX_01	EGR1_01.ER_Q6	MZF1_01.P53_01
AP2_Q6.MZF1_01	EGR1_01.HNF4_01	**MZF1_01.SP1_Q6**
AP2_Q6.NFKAPPAB_01	EGR1_01.MYOD_Q6	NF1_Q6.NRSF_01
AP2_Q6.P53_01	EGR1_01.MZF1_01	NRSF_01.P300_01
AP2_Q6.SREBP1_02	**EGR1_01.NF1_Q6**	NRSF_01.RREB1_01
AREB6_03.ARNT_01	**EGR1_01.P300_01**	NRSF_01.SP1_Q6
AREB6_03.EGR1_01	EGR1_01.PAX2_01	NRSF_01.PAX2_01
AREB6_03.NRSF_01	EGR1_01.PAX5_01	NRSF_01.PAX5_01
AREB6_03.P53_01	EGR1_01.RREB1_01	NRSF_01.YY1_02
AREB6_03.SREBP1_02	**EGR1_01.SP1_Q6**	**P53_01.SP1_Q6**
ARNT_01.ER_Q6	EGR1_01.YY1_02	**SP1_Q6.SREBP1_02**
ARNT_01.ELK1_02	ELK1_02.NRSF_01	**SP1_Q6.USF_01**
ARNT_01.MYOD_Q6	ER_Q6.NRSF_01	
ARNT_01.MZF1_01	MAX_01.MYOD_Q6	
ARNT_01.P300_01	MAX_01.MZF1_01	

Promoter		
AREB6_03.EGR1_01	CREBP1_01.SRY_02	FOXD3_01.MEF2_03
AREB6_03.NRSF_01	E2F_02.MZF1_01	FREAC3_01.SRY_02
CART1_01.FOXD3_01	E2F_02.NRSF_01	FREAC4_01.GATA_C
CART1_01.SRY_02	**EGR1_01.ELK1_02**	FREAC4_01.SRY_02
CDP_02.FOXD3_01	EGR1_01.MZF1_01	HNF1_01.TST1_01
CDP_02.GATA_C	**EGR1_01.P300_01**	IRF1_01.SRY_02
CDP_02.SOX9_B1	EGR1_01.PAX2_01	MEF2_03.SRY_02
CDP_02.SRY_02	EGR1_01.PAX5_01	MZF1_01.NRSF_01
CREBP1_01.FOXD3_01	EGR1_01.YY1_02	PBX1_02.SRY_02

Mouse		
AML1_01.ARNT_01	ATF_01.NFE2_01	**MAX_01.SP1_Q6**
AML1_01.ATF_01	ATF_01.P300_01	MEF2_03.NKX61_01
AML1_01.E2F_02	ATF_01.PAX2_01	MEF2_04.NKX61_01
AML1_01.EGR1_01	CDP_02.FOXD3_01	MEIS1_01.NFKAPPAB_01
AML1_01.MYCMAX_01	CDP_02.FOXJ2_02	MEIS1_01.NRSF_01
AML1_01.NFKAPPAB_01	CDP_02.NKX61_01	MEIS1_01.SP1_Q6
AML1_01.NRSF_01	CREBP1_01.FOXD3_01	MEIS1_01.SREBP1_02
AML1_01.P300_01	CREBP1_01.FOXJ2_02	MYB_Q6.NFKAPPAB_01
AML1_01.SP1_Q6	CREBP1_01.NKX61_01	MYB_Q6.NRSF_01
AML1_01.SREBP1_02	CREL_01.E2F_02	MYB_Q6.SP1_Q6
AML1_01.USF_01	CREL_01.EGR1_01	MYB_Q6.SREBP1_02
AML1_01.YY1_02	CREL_01.MZF1_01	MYCMAX_01.MYOD_Q6
AP1_Q2.NFKAPPAB_01	CREL_01.NRSF_01	MYCMAX_01.MZF1_01
AP2_Q6.ARNT_01	CREL_01.SREBP1_02	MYCMAX_01.NF1_Q6
AP2_Q6.ATF_01	E2F_02.ELK1_02	MYCMAX_01.NFE2_01
AP2_Q6.E2F_02	E2F_02.ER_Q6	MYCMAX_01.P300_01
AP2_Q6.E47_02	E2F_02.GATA2_01	MYCMAX_01.PAX2_01
AP2_Q6.EGR1_01	E2F_02.MYB_Q6	MYCMAX_01.PAX5_01
AP2_Q6.GATA2_01	E2F_02.MYOD_Q6	**MYCMAX_01.SP1_Q6**
AP2_Q6.MAX_01	E2F_02.MZF1_01	MYCMAX_01.YY1_02
AP2_Q6.MYCMAX_01	E2F_02.NF1_Q6	MYOD_Q6.NFE2_01
AP2_Q6.MZF1_01	E2F_02.P300_01	MYOD_Q6.NFKAPPAB_01
AP2_Q6.NFKAPPAB_01	E47_02.MZF1_01	MYOD_Q6.NRSF_01
AP2_Q6.NRSF_01	E47_02.P300_01	MYOD_Q6.PAX5_01
AP2_Q6.P300_01	**EGR1_01.ELK1_02**	**MYOD_Q6.SP1_Q6**
AP2_Q6.P53_01	EGR1_01.ER_Q6	MYOD_Q6.SREBP1_02
AP2_Q6.SREBP1_02	EGR1_01.GATA2_01	MYOD_Q6.USF_01
AP2_Q6.USF_01	EGR1_01.HNF4_01	MYOD_Q6.YY1_02
AREB6_03.ARNT_01	EGR1_01.MYOD_Q6	MZF1_01.NFE2_01
AREB6_03.ATF_01	EGR1_01.MZF1_01	MZF1_01.NFKAPPAB_01
AREB6_03.E2F_02	**EGR1_01.NF1_Q6**	MZF1_01.NRSF_01
AREB6_03.EGR1_01	**EGR1_01.P300_01**	MZF1_01.P53_01
AREB6_03.GATA2_01	EGR1_01.PAX2_01	MZF1_01.PAX2_01
AREB6_03.MYCMAX_01	EGR1_01.PAX5_01	MZF1_01.PAX5_01
AREB6_03.MYOD_Q6	EGR1_01.RREB1_01	**MZF1_01.SP1_Q6**
AREB6_03.MZF1_01	**EGR1_01.SP1_Q6**	MZF1_01.SREBP1_02
AREB6_03.NF1_Q6	EGR1_01.YY1_02	MZF1_01.USF_01
AREB6_03.NRSF_01	ELK1_02.MYCMAX_01	MZF1_01.YY1_02
AREB6_03.P300_01	ELK1_02.MZF1_01	**NF1_Q6.NFKAPPAB_01**
AREB6_03.SREBP1_02	**ELK1_02.NFKAPPAB_01**	NF1_Q6.NRSF_01
AREB6_03.USF_01	ELK1_02.NRSF_01	**NF1_Q6.SP1_Q6**
ARNT_01.ELK1_02	**ELK1_02.P300_01**	NF1_Q6.SREBP1_02
ARNT_01.MYOD_Q6	ELK1_02.SREBP1_02	**NFKAPPAB_01.P300_01**
ARNT_01.MZF1_01	ER_Q6.MYCMAX_01	NKX61_01.SRY_02
ARNT_01.P300_01	ER_Q6.MZF1_01	NRSF_01.P300_01
ARNT_01.PAX5_01	ER_Q6.NFKAPPAB_01	NRSF_01.PAX2_01

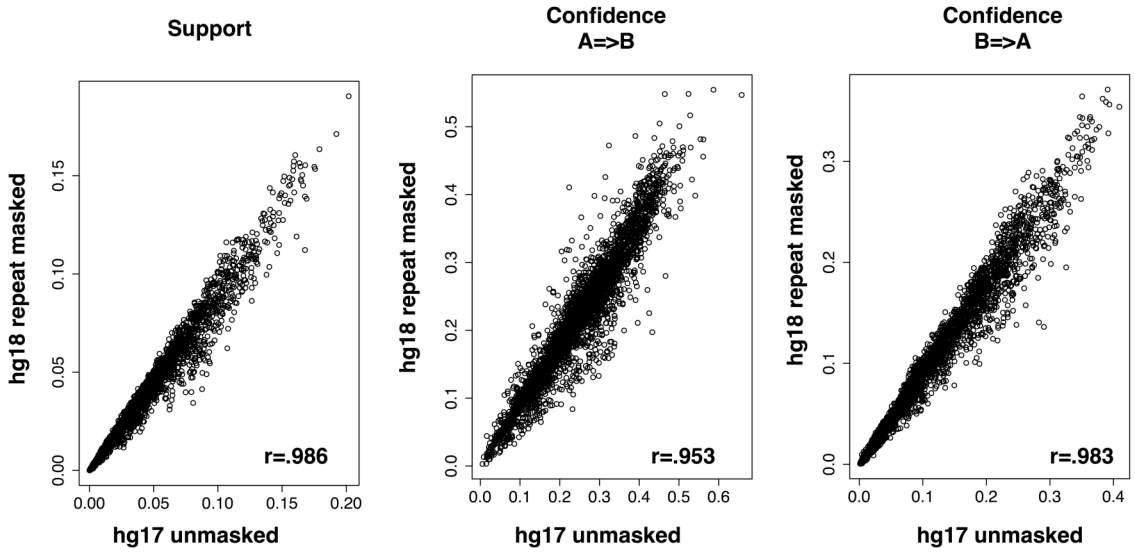
Mouse		
ARNT_01.SP1_Q6	ER_Q6.NRSF_01	NRSF_01.PAX5_01
ARNT_01.YY1_02	FOXD3_01.HNF1_01	NRSF_01.RREB1_01
ARP1_01.ATF_01	FOXD3_01.NKX61_01	NRSF_01.SP1_Q6
ARP1_01.EGR1_01	FOXJ2_02.HLF_01	NRSF_01.YY1_02
ARP1_01.MYOD_Q6	FOXJ2_02.HNF1_01	**P300_01.PAX5_01**
ARP1_01.MZF1_01	GATA2_01.MYCMAX_01	**P300_01.SP1_Q6**
ARP1_01.NRSF_01	**GATA2_01.NFKAPPAB_01**	P300_01.SREBP1_02
ARP1_01.SREBP1_02	GATA2_01.NRSF_01	P300_01.USF_01
ATF_01.CREL_01	**GATA2_01.SP1_Q6**	**P300_01.YY1_02**
ATF_01.ELK1_02	GATA2_01.SREBP1_02	**P53_01.SP1_Q6**
ATF_01.ER_Q6	GATA2_01.USF_01	**SP1_Q6.USF_01**
ATF_01.HNF4_01	HNF4_01.NFKAPPAB_01	
ATF_01.MEIS1_01	HNF4_01.NRSF_01	
ATF_01.MYB_Q6	HSF2_01.MZF1_01	
ATF_01.MYOD_Q6	MAX_01.MYOD_Q6	
ATF_01.MZF1_01	MAX_01.MZF1_01	
ATF_01.NF1_Q6	MAX_01.P300_01	

Supplemental Table 2.3: Estimated rates of Patser error

The estimated rates of Patser overestimation for PWMs.

TF	Est. Error		TF	Est. Error
AML1_01	0.19		LMO2COM_02	0.40
AP1_Q2	0.18		MAX_01	0.24
AP2_Q6	0.07		MEF2_03	0.18
AP4_Q6	0.11		MEF2_04	0.12
AREB6_03	0.10		MEIS1_01	0.12
ARNT_01	0.30		MIF1_01	0.17
ARP1_01	0.11		MYB_Q6	0.24
ATF_01	0.41		MYCMAX_01	0.19
BRN2_01	0.21		MYOD_Q6	0.09
CART1_01	0.26		MZF1_01	0.08
CDP_02	0.54		NF1_Q6	0.10
CEBP_C	0.20		NFAT_Q6	0.09
CHOP_01	0.21		NFE2_01	0.15
COUP_01	0.12		NFKAPPAB_01	0.11
CREBP1_01	0.61		NFY_01	0.25
CREB_01	0.51		NKX61_01	0.15
CREL_01	0.15		NRSF_01	0.04
E2F_01	0.44		OCT_C	0.16
E2F_02	0.52		P300_01	0.11
E47_02	0.09		P53_01	0.09
EGR1_01	0.17		PAX2_01	0.29
ELF1_01	0.23		PAX5_01	0.05
ELK1_02	0.18		PAX6_01	0.28
ER_Q6	0.16		PBX1_02	0.24
FOXD3_01	0.12		RFX1_01	0.22
FOXJ2_02	0.17		RORA1_01	0.18
FREAC2_01	0.21		RORA2_01	0.20
FREAC3_01	0.19		RREB1_01	0.04
FREAC4_01	0.17		SOX9_B1	0.20
FREAC7_01	0.15		SP1_Q6	0.04
GATA2_01	0.26		SREBP1_02	0.13
GATA3_01	0.22		SRF_Q6	0.17
GATA_C	0.23		SRY_02	0.17
GRE_C	0.13		STAT_01	0.17
HFH3_01	0.12		TATA_01	0.28
HLF_01	0.38		TCF11_01	0.20
HNF1_01	0.22		TGIF_01	0.18
HNF4_01	0.1		TST1_01	0.24

TF	Est. Error		TF	Est. Error
HSF1_01	0.25		USF_01	0.19
HSF2_01	0.27		XBP1_01	0.48
IRF1_01	0.12		YY1_02	0.1
ISRE_01	0.09			



Supplemental Figure 2.1: Effects of repeat masking

Support, confidence A=>B, and confidence B=>A are highly correlated between hg17 without repeat masking and hg18 with repeat masking.

Chapter 3: Materials and Methods

MICROARRAYS

Generating Spotted cDNA arrays

Spotted cDNA arrays were made as previously described [15, 152]. Briefly, to generate the yeast arrays used in the Iyer lab, each ORF and intergenic region in the yeast genome was amplified by large-scale PCR using PCR primers from the Research Genetics yeast ORF library. The template for PCR was genomic DNA from the yeast strain S288C. Amplified DNA was precipitated was precipitated with acid (10% volume 3 M sodium acetate) and 2.5 volumes 95% ethanol, purified, transferred to 384-well plates, and dried. Before printing, DNA was rehydrated with 6 μL 3x SSC solution added by a BioMek robot. A custom-built robot arrayer with 48 hollow-tip pins was used to pick up DNA, spot it onto polylysine-coated slides, and wash and dry the tips after each load.

Polylysine slides

To create polylysine slides, microscope slides were washed for 2 hours in a solution of 206 g NaOH in 2 L ddH20 on a gyratory shaker, then rinsed thoroughly with ddH2O and drained. Slides were shaken for 40 minutes in polylysine solution (750 mL ddH2O, 105 mL PBS, 150 mL poly-L-lysine), rinsed in ddH2O for 5 minutes, and dried by centrifugation for 5 minutes at 42 G. Slides were stored in a plastic box and aged until they reached proper hydrophobicity (~2 weeks).

Post-processing

After printing and before use, spotted cDNA arrays must be post-processed. Post processing denatures DNA to allow hybridization of probe and caps exposed amines in

the polylysine so that the probe can only bind to the DNA [9, 153]. Succinic anhydride is used to cap amines, and boiling is used to denature DNA. To post-process, the backs of slides were etched with an etching pen to note the location of the spots. 5.5 g succinic anhydride was dissolved in 335 mL 1-methyl-2-pyrrolidinone. 15 mL of 1 M sodium borate (pH 8.0) was added immediately after dissolution of the succinic anhydride. Slides were added to the solution and vigorously mixed for 30 seconds; slides in solution were then shaken for 15 minutes on a gyratory shaker. Next, slides were placed in 95°C ddH2O for 90 seconds, then transferred to 95% ethanol. Slides were dried by centrifugation at 42 G for 5 minutes in a tabletop centrifuge and used immediately.

YEAST CULTURE

Yeast from freezer stocks was streaked on YPD plates and grown for three days at 30°C. When growing cultures for total RNA isolation or for chromatin immunoprecipitation, a single yeast colony from a plate was placed in 2 mL YPD and shaken at 200 RPM overnight at 30°C. This culture was then spun down at 3200 G, resuspended in 1 M sorbitol, and then added to synthetic complete media for an OD 600 nm of ~0.1. This culture was shaken at 200 RPM at 30°C until OD 600 nm ~0.5. Methyl methane sulfonate (Sigma) was added to MMS-treated cultures for a total concentration of .02%, and cultures were harvested after one hour. Cultures for chromatin immunoprecipitation [153] were cross-linked by adding formaldehyde (Merck) for a final concentration of 1% for 30 minutes, then quenched with 2.5 M glycine for a final concentration of 125 mM for ten minutes before collection. Cells for total RNA isolation were harvested by centrifugation at 3200 G for 10 minutes. Cells for chromatin immunoprecipitation were harvested by centrifugation at 3200 G for 5 minutes at 4°C and washed twice with PBS. All cell pellets were then stored at -80°C.

TOTAL RNA ISOLATION

Total RNA was isolated from yeast as described by DeRisi [154]. Briefly, yeast pellets were resuspended in 8 mL AE buffer (50mM sodium acetate pH 5.2, 10 mM EDTA). Eight mL acid phenol (pH 4.5 – 5.5) (Anachemia) was added, as was SDS (sodium dodecyl sulfate) to a final concentration of 0.8%. This mixture was then incubated at 65°C for 1 hour with vortexing every 15 minutes. Next, the lysate was incubated on ice for 10 minutes before centrifugation at 3200 G for 15 minutes to pellet cell debris. The supernatant was transferred to a fresh tube, combined with 8 mL acid phenol, vortexed, and centrifuged again at 3200 G for 15 minutes. The supernatant was then transferred to a phase lock tube and 10 mL chloroform was added. The tube was shaken thoroughly and centrifuged at 1833 G for 10 minutes before transferring the supernatant to a fresh tube. RNA was precipitated with acid (10% volume 3 M sodium acetate) and 2.5 volumes 95% ethanol, and the tubes were spun for 40 minutes at 3200 G. The RNA pellet was washed three times with 70% ethanol, dried by vacuum, resuspended in DEPC water, and stored at -80°C. RNA concentration was quantified by Nanodrop, and quality was assessed by 1% agarose gel.

REVERSE TRANSCRIPTION

Reverse transcription allows incorporation of 5' amino allyl dUTP into the cDNA strand to allow coupling of Cy3 and Cy5 fluorescent dyes. For reverse transcription, 15 μg total RNA was placed in a PCR tube for a total volume of 15 μL RNA + DEPC water. One μL of oligo dT primer (5'-TTT TTT TTT TTT TTT TTT TTV N-3') (5 μg/μL) was added. The mixture was incubated at 70°C for 10 minutes, chilled on ice for 10 minutes, and a reaction mix was added consisting of 6 μL 5x 1st strand buffer (Invitrogen), 3 μL 0.1 M DTT, 1.2 μL 25x amino allyl dUTP/dNTP mix, 3 μL H20, and 2 μL Superscript II (Invitrogen). The sample was incubated for 2 hours at 42°C, raised to 95°C for 5

minutes, then snapped cool on ice for 5 minutes. RNA was then degraded by adding 13 μL 1 M NaOH and 1 μL 0.5 M EDTA. The mixture was incubated for an additional 15 minutes at 67°C and then neutralized with 50 μL 1 M HEPES buffer (pH 7.5). The pH was decreased to ~7.0 with the addition of 35 μL 3 M sodium acetate (pH 5.2) to allow for cleanup with a Qiagen MinElute kit, which was used according to the manufacturer's protocol. Elution was with 0.1 M sodium bicarbonate (pH 9.0).

cDNA Labeling

Cy3 and Cy5 dye packs (Amersham) were resuspended in 3.6 μL DMSO. 1.2 μL dye was added to each sample, which was then incubated at room temperature for one hour. The sample was then cleaned with a Qiagen MinElute kit according to the manufacturer's instructions.

Hybridization

Each Cy3-labeled sample was combined with its respective Cy5-labeled sample as well as 1 μL tRNA (5 μg/μL), 1 μL polyA (10 μg/μL), and 1 μL 1 M HEPES buffer (pH 7.5). The final sample volume was adjusted to contain 17.5% 20X SSC and 0.25% SDS. The sample was incubated at 100°C for 2 minutes to denature DNA, spun down, and applied to the array. The array was incubated in a 65°C water bath for 6-16 hours.

Array Washing

Arrays were first washed for 30 seconds in 350 mL of 57% SSC and 0.028% SDS solution. They were then washed in 350 mL of 5.7% SSC solution before drying by centrifugation at 42 G for 1 minute.

Array Scanning, normalization, and analysis

Arrays were scanned with an Axon 4000b scanner using GenePix 5.1, gridded, and uploaded to and normalized by the Longhorn Microarray Database [155, 156].

Error Model

Background

The microarray expression changes that are most believable are those that are highly repeatable between biological replicates. It is useful to average biological replicates, but differences in array quality may mean that some arrays should be weighted more heavily than others. Here we describe an error model that accounts for the experimental variance of each array when combining biological replicates. This error model was adapted in our lab [157, 158] from the model previously described by Hughes [159] and similar to that used by Winzeler [160]. In this model, error is estimated from 10 same-versus-same control experiments in which identical RNA samples are labeled with Cy3 and Cy5 dyes and hybridized to arrays. The distribution of intensity ratios for these controls reveals the error distribution: higher-intensity log ratios, which correspond to abundant transcripts, are quite constant from one experiment to another, while lower-intensity ratios, which correspond to low transcript levels, are much more variable. The error model has an additive normal intensity component σ, which corresponds to error caused by background subtraction, and a multiplicative component *f*, which corresponds to error due to scanner fluctuations, lack of uniformity in labeling, or divergent hybridization efficiencies.

Calculation

First, a significance score X is assigned to a spot ratio using the following equation:

$$X = \frac{R - G}{\sqrt{\sigma_1^2 + \sigma_2^2 + f^2(R^2 + G^2)}} \qquad \text{(A)}$$

Here R and G are the normalized net intensities of the array channels. σ_1^2 and σ_2^2 model background subtraction error, which is highly variable and is represented by the standard deviation of local background in each channel [159]. The multiplicative error component f, which is similar between arrays, is calculated from the control experiments as [157, 158]:

$$f = stdev(\ln(\frac{R}{G})) \qquad \text{(B)}$$

Because X is normally distributed, the *P*-value of observing a spot with significance |X| is equal to:

$$p = 2 \times normcdf(-|X|) \qquad \text{(C)}$$

Expressing intensities as log ratios facilitates averaging of array replicates, but it requires correction of intensities that became negative due to background subtraction. In our analysis, if both channels had negative intensity, the spot was omitted from analysis. If only one channel was negative, intensity equal to local background standard deviation was assigned. If this resulted in a higher intensity than the net intensity of the other channel, the spot was omitted from analysis. Due to the subjective nature of log transformation, a measure of log ratio uncertainty was assigned to each spot:

$$\sigma_{\log_{10}\left(\frac{R}{G}\right)} = \frac{\log_{10}\left(\frac{R}{G}\right)}{X} \qquad \text{(D)}$$

This ratio allows biological replicates to be combined and appropriately weighted. As a result, dim spots have small X scores and large uncertainty, while bright spots have larger X scores and smaller uncertainty. For

each array spot, mean $\log_{10}$(R/G) is calculated by minimum variance-weighted average, where σ_i^2 is the error of the log ratio and X_i is the *i*-th replicate of the ratio[159]:

$$w = \frac{1}{\sigma_i^2} \qquad \text{(E)}$$

$$\overline{X} = \frac{\sum_{i=1,n} w_i x_i}{\sum_{i=1,n} w_i} \qquad \text{(F)}$$

The significance of $\overline{X}$ was calculated using equation C.

Implementation

Our use of the error model [157, 158] differed slightly from that of other studies [17, 160]. Although all used the minimum variance weighted average method (D) $\log_{10}$(R/G), one [160] proceeded to estimate error for the mean ratio to derive an X score for which a *P*-value was derived according to equation C. Our lab found [157, 158] that this approach was so highly dependent on the repeatability of biological replicate measures that small but repeatable gene changes of dubious biological meaning were overly represented in significant results. Another study [17] standardized $\overline{X}$ by Z-score transformation; this presupposed that each experiment would yield identical ratio distributions and numbers of target genes, but the deletion of individual transcription factors can affect the expression of hundreds of genes or only a few [158].

To calculate differentially expressed genes in each deletion experiment, microarray data was extracted from the Longhorn Microarray database [158]. R and G

were represented by normalized channel 2 (Cy5) median net intensity and channel 1 (Cy3) median net intensity, respectively. σ_1 and σ_2 were represented by the standard deviation of channel 1 and channel 2 background, respectively. A Perl script was used to calculate X scores for each spot, and custom Java software [158] was used to correct negative spots, to determine log ratio uncertainties, and to calculate weighted mean $\overline{X}$ for each gene. A significance cutoff of $P < 0.05$ was used for chromatin immunoprecipitation data, and a cutoff of $P < 0.001$ was used for expression data.

MEDIA COMPOSITION

Difco YPD was used for all YPD media and plates. YPD plates also include 2% agar. Synthetic complete media consisted of 2% glucose, 1X yeast nitrogen base (Difco), 1x SD –ura dropout solution (Difco), and 25 μg/mL uracil.

25X amino allyl dUTP/dNTP mix was composed of 41.7 μL each 100 mM dATP, dGTP, and dCTP, 16.7 μL 100 mM dTTP, 50 μL 5' amino-allyl dUTP (Ambion), and 141.7 μL ddH2O.

CHROMATIN IMMUNOPRECIPITATION FOR TAP-TAGGED STRAINS

Cell pellets were resuspended in 2 mL cold lysis buffer (50 mM HEPES-KOH pH 7.5, 150 mM KCl, 1 mM EDTA, 10% glycerol, 0.1% NP40, and protease inhibitor (Roche)), then disrupted by bead-beating (Biospec Mini-BeadBeater-8) for five one-minute sessions, with incubation on ice for 2 minutes between each session. Cell lysate was sonicated for 30 minutes with a Bioruptor (Diagenode) to shear the DNA and centrifuged to pellet debris at 16,000 G for 10 minutes at 4°C. DNA fragment size (300 – 1000 bp) was verified by agarose gel. 500 μL cell lysate was pre-cleared with 25 μL protein G agarose beads (Roche) for 1 hour at 4°C, then incubated overnight at 4°C with 2 μg anti-PAP. Lysate was then incubated for 2 hours at 4°C with protein G agarose

beads. Beads were washed twice for five minutes with IP wash buffer 1 (50 mM HEPES-KOH pH 7.5, 50 mM NaCl, 1 mM EDTA, 0.1% sodium deoxycholate, and 1% Triton X-100), IP Wash Buffer 2 (50 mM HEPES-KOH pH 7.5, 500 mM NaCl, 1 mM EDTA, 0.1% sodium deoxycholate, and 1% Triton X-100), and IP Wash Buffer 3 (10 mM TRIS-HCl pH 8.0, 250 mM LiCl, 1 mM EDTA, 0.5% sodium deoxycholate, 0.5% NP-40), respectively. Beads were then washed once with TE (10 mM Tris-HCl pH 8.0, 1 mM EDTA) for five minutes and eluted twice by addition of 100 μL elution buffer (50 mM TRIS-HCl pH 8.0, 10 mM EDTA, 1% SDS), incubation at 65°C for 30 minutes, and centrifugation at 12,000 G for 2 minutes. Crosslinks were reversed by heating overnight at 65°C. The sample was incubated for two hours at 37°C following addition of 150 μL TE, 1 μL glycogen (20 mg/mL), and Proteinase K to a final concentration of 100 μg/mL. The sample was extracted twice with 300 μL phenol: chloroform (Invitrogen) before addition of 1.5 μL RNAse A (1 μg/μL) and incubation for 30 minutes at 37°C. DNA was precipitated by addition of sodium acetate (pH 5.2) to a final concentration of 0.3 M and two volumes of 95% ethanol, pelleted, washed with 70% ethanol, dried, resuspended in TE, and stored at -20°C.

ROUND A-B PCR AMPLIFICATION

Background

The yield of DNA from chromatin immunoprecipitation (~30 ng) is insufficient for microarray hybridization. Therefore, a two-round PCR-based strategy [15] was used to amplify immunoprecipitated DNA. In the first round, Primer A (assorted random primers with a known 5' end) was annealed to the denatured DNA and extended with *exo-Klenow*. This round was repeated so that each immunoprecipitated DNA fragment had a corresponding fragment with two known ends corresponding to the 5' end of

Primer A, while the original proportions of DNA in the pool were preserved. In the second round of amplification, Round A product was used as a template, and Primer B (corresponding to the known 5' end of Primer A) was used to amplify this template by polymerase chain reaction. Amino allyl dUTP is used in the PCR to allow coupling of Cy3 and Cy5 dyes.

Round A Protocol

8 μL DNA from chromatin immunoprecipitation (2.5 – 100 ng /μL), 1 μL 10X Klenow buffer (USB), and 1 μL Primer A (40 μM, sequence 5' GTTTCCCAGTCACGATCNNNNNNNNN 3') were added to an MJ Research thermal cycler and incubated for 2 minutes at 94°C. Temperature was reduced to 8°C, and 0.5 μL 10X Klenow buffer, 1.5 μL 5mM dNTPs, 0.5 μL *exo-Klenow* (2 U/μL), and 2.5 μL ddH20 were added to each reaction. Incubation temperature was increased from 8°C to 37°C over a period of eight minutes. The mixture was incubated at 37°C for 40 minutes, then at 95°C for 5 minutes. Temperature was then decreased to 8°C, and 0.5 μL *exo-Klenow* was added to each reaction. Temperature was increased to 37°C over a period of eight minutes, where the reaction was incubated for a further 40 minutes. 85 μL ddH20 was then added to each reaction.

Round B Protocol

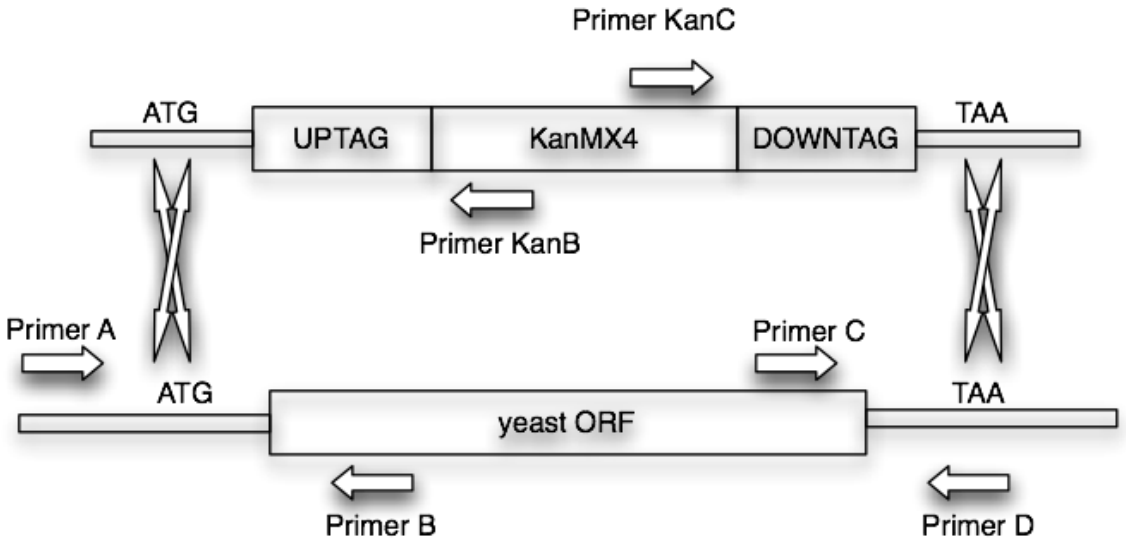
Each reaction consisted of 15 μL Round A product, 8 μL magnesium chloride (25 mM), 10 μL 10X PCR buffer (Applied Biosystems), 2 μL 25X aa-dUTP/dNTPs, 2.5 μL Primer B (500 μM, sequence 5' GTTTCCCAGTCACGATC 3'), 1 μL Taq polymerase (5 U/μL), and 61.5 μL ddH20. The PCR program was 32 cycles of (92°C for 30 seconds, 40°C for 30 seconds, 50°C for 30 seconds, 72°C for 60 seconds). Quality was assessed

by 1% agarose gel, and PCR product was cleaned with a Qiagen MinElute kit according to the manufacturer's instructions.

VERIFICATION OF KNOCKOUTS BY PCR

Background

All strains were derived from the Saccharomyces Genome Deletion Project library, which was purchased from Open Biosystems. These strains are created using a PCR-based gene deletion strategy [161, 162], in which homologous recombination is used to replace each gene with a KanMX4 cassette tagged with one or two unique 20-mer sequences (Figure 3.1). This facilitates simple verification of knockouts by PCR; each knockout will have a PCR band of known length for primers A and KanB, and for primers D and KanC. Primers A and D will yield a specific PCR product of one length for a knockout and another for a wild type. In our studies, genomic DNA was isolated from knockouts, and primers A and KanB were used for PCR.



Adapted from Saccharomyces Genome Deletion Project
http://www-sequence.stanford.edu/group/yeast_deletion_project/confirmation.html

Figure 3.1: Yeast deletion PCR strategy

Primers A, B, C, and D are specific for each gene. The ORF is replaced with the KanMX4 kanamycin resistance cassette via homologous recombination. Successful knockouts are confirmed by PCR with primers A and KanB, or with primers D and KanC.

ISOLATION OF GENOMIC DNA

Genomic DNA was extracted as described by Hoffman [163]. Briefly, a single colony of yeast was placed in 5 mL YPD and incubated 24 hours at 30°C with shaking at 200 RPM. Yeast was pelleted by centrifugation for 5 minutes at 1200 G, then resuspended in 0.5 mL sorbitol solution (0.9 M sorbitol, 0.1 M Tris-HCl pH 8.0, 0.1 M EDTA). Spheroplasts were prepared by addition of 50 μL zymolase solution (0.3 mg/mL) and 50 μL 0.28 M β-mercaptoethanol, followed by incubation for 1 hour at 37°C and 200 RPM. Cells were pelleted by centrifugation for 5 minutes at 1200 G, then resuspended in 0.5 mL Tris/EDTA solution (50 mM Tris-HCl pH 8.0, 20 mM EDTA). Cells were lysed by addition of 50 μL 10% SDS and incubated for 20 minutes at 65°C. 200 μL 5 M potassium acetate was added to the cells; they were incubated on ice for 30 minutes, then centrifuged for 3 minutes at 16,000 G. The supernatant was precipitated by addition of 1mL 100% ethanol. Pelleted DNA was dried for 5 minutes in a Speedvac evaporator and resuspended in 300 μL TE. DNA was then incubated for 1 hour at 37°C with 5 μL RNAse A (1 mg/mL). DNA was re-precipitated with 500 μL isopropanol, removed from the solution with a 200 μL pipette tip, squeezed dry against the wall of the tube, and resuspended in 125 μL TE.

VERIFICATION OF GENE DELETION BY PCR

Each reaction consisted of 5 μL 10X PCR buffer (Sigma), 4 μL magnesium chloride (25 mM), 0.5 μL dNTPs (25 mM), 5 μL of each primer (10 μM), 60 ng genomic DNA isolated as described above, 2 μL Taq polymerase (5 U/μL), and 28 μL ddH20. The PCR program was 94°C for 3 minutes, followed by 35 cycles of (94°C for 15 seconds, 57°C for 15 seconds, 72°C for 1 minute), and ending with 72°C for 5 minutes. Bands were assessed on 1% agarose gel.

YEAST TRANSFORMATION

A 25 mL overnight culture of wild-type yeast was grown to OD 600 nm ~0.5, then pelleted by centrifugation at 1833 G for 2 minutes. The pellet was washed once in 1 mL ddH2O, then resuspended in 1 mL 100 mM lithium acetate and incubated for 5 minutes at 30°C. For each transformation reaction, 100 μL of cells in lithium acetate were spun down in a microcentrifuge, and supernatant was removed. 240 μL polyethylene glycol (50% weight by volume), 36 μL 1 M lithium acetate, 8 μL denatured, sheared salmon sperm DNA, 5 μL transformation DNA (100 ng to 5 μg total), and 65 μL ddH2O were added to the cells. Cells were then vortexed for one minute and incubated for 20 minutes at 42°C. Cells were pelleted in a microcentrifuge, resuspended in 1 mL YPD, and incubated for 2.5 hours at 50 RPM. Cells were then pelleted, resuspended in 400 μL YPD, and spread on plates with selective media. Plates were incubated for 2-4 days at 30°C.

Table 3.1: Yeast strains used in this study

ORF	Name	Mother Strain	Genotype
YDR216W	ΔAdr1	BY4741	MATa his3Δ1 leu2Δ0 met15Δ0 ura3Δ0, adr1Δ
YJL115W	ΔAsf1	BY4741	MATa his3Δ1 leu2Δ0 met15Δ0 ura3Δ0, asf1Δ
YOR113W	ΔAzf1	BY4741	MATa his3Δ1 leu2Δ0 met15Δ0 ura3Δ0, azf1Δ
YDR423C	ΔCad1	BY4741	MATa his3Δ1 leu2Δ0 met15Δ0 ura3Δ0, cad1Δ
YNL068C	ΔFkh2	BY4741	MATa his3Δ1 leu2Δ0 met15Δ0 ura3Δ0, fkh2Δ
YLR013W	ΔGat3	BY4741	MATa his3Δ1 leu2Δ0 met15Δ0 ura3Δ0, gat3Δ
YOL012C	ΔHtz1	BY4741	MATa his3Δ1 leu2Δ0 met15Δ0 ura3Δ0, htz1Δ
YLR384C	ΔIki3	BY4741	MATa his3Δ1 leu2Δ0 met15Δ0 ura3Δ0, iki3Δ
YOR123C	ΔLeo1	BY4741	MATa his3Δ1 leu2Δ0 met15Δ0 ura3Δ0, leo1Δ
YNR052C	ΔPop2	BY4741	MATa his3Δ1 leu2Δ0 met15Δ0 ura3Δ0, pop2Δ
YNL250W	ΔRad50	BY4741	MATa his3Δ1 leu2Δ0 met15Δ0 ura3Δ0, rad50Δ
YER095W	ΔRad51	BY4741	MATa his3Δ1 leu2Δ0 met15Δ0 ura3Δ0, rad51Δ
YML032C	ΔRad52	BY4741	MATa his3Δ1 leu2Δ0 met15Δ0 ura3Δ0, rad52Δ
YLR039C	ΔRic1	BY4741	MATa his3Δ1 leu2Δ0 met15Δ0 ura3Δ0, ric1Δ
YHR154W	ΔRtt107	BY4741	MATa his3Δ1 leu2Δ0 met15Δ0 ura3Δ0, rtt107Δ
YMR190C	ΔSgs1	BY4741	MATa his3Δ1 leu2Δ0 met15Δ0 ura3Δ0, sgs1Δ
YML081W	ΔYML081W	BY4741	MATa his3Δ1 leu2Δ0 met15Δ0 ura3Δ0, yml081wΔ
YDR369C	ΔXrs2	BY4741	MATa his3Δ1 leu2Δ0 met15Δ0 ura3Δ0, xrs2Δ
YDR216W	Adr1-TAP	BY4741	MATa his3D1 leu2D0 met15D0 ura3D0 adr1-TAP::HIS3MX6
YOR113W	Azf1-TAP	BY4741	MATa his3D1 leu2D0 met15D0 ura3D0 azf1-TAP::HIS3MX6
YLR384C	Iki3-TAP	BY4741	MATa his3D1 leu2D0 met15D0 ura3D0 iki3-TAP::HIS3MX6

ORF	Name	Mother Strain	Genotype
YLR039C	Ric1-TAP	BY4741	MATa his3D1 leu2D0 met15D0 ura3D0 ric1-TAP::HIS3MX6
YHR154W	Rtt107-TAP	BY4741	MATa his3D1 leu2D0 met15D0 ura3D0 rtt107-TAP::HIS3MX6
YML081W	YML081W-TAP	BY4741	MATa his3D1 leu2D0 met15D0 ura3D0 yml081w-TAP::HIS3MX6
	Wild Type	BY4741	MATa his3D1 leu2D0 met15D0 ura3D0

Table 3.2: Primers used for deletion verification

Primer Name	**Primer Sequence (5'-3')**
ΔAdr1 Primer A	CATTGATCTGAATTTCTCAGGCTAT
ΔAsf1 Primer A	AATGCTGTTTTATTCCGTTCTTACA
ΔAzf1 Primer A	ATCCCAAGACTTATATAGCCCTACG
ΔCad1 Primer A	CAAGTACCAGTTGGAAAGAGACATT
ΔFkh2 Primer A	ATCTTCGATTTCGCTCATTAAAAG
ΔGat3 Primer A	GTATTTCTGATAAAATGCGGACAAC
ΔHtz1 Primer A	TCCATGCTAGATTAGCACACAGTAA
ΔIki3 Primer A	CTGTTAAAAGATCCCGTCATTGATA
ΔLeo1 Primer A	ACAAGGACAAGAAGGTGATATTGAG
ΔPop2 Primer A	TGTTCTTATTATATGGCAGCAAACA
ΔRad50 Primer A	ATAACCATGCATCTTGCAATACTTT
ΔRad51 Primer A	CCAATCTAGTTTAGCTATCCTGCAA
ΔRad52 Primer A	GATTCAACAACTCCCTTGGCGTC
ΔRic1 Primer A	CAACCAAATTTGACAATTTAATTCC
ΔRtt107 Primer A	ACTTAACCACAGAATGTTCTTCGAC
ΔSgs1 Primer A	CCTGATCTAAAAGCTGATATACGGA
ΔYML081W Primer A	TAATCTTTTTTTTTGCTGAAAAACCC
ΔYML081W Primer D	TTGACGATCACCTTGCTTGTTTTTATT
ΔXrs2 Primer A	GTATTGAAGCAATTTGTAAGCTGGT
KanB	CTGCAGCGAGGAGCCGTAAT
KanC	TGATTTTGATGACGAGCGTAAT

Chapter 4: Transcriptional profiling of MMS-sensitive yeast mutants

ABSTRACT

Repair of double-stranded DNA breaks is a highly conserved process essential to cell survival. To identify novel genes involved in double-stranded break repair in yeast, we screened ~350 transcription factors and DNA-binding proteins for sensitivity to the cross-linking agent methyl methane sulfonate, then performed transcriptional profiling and chromatin immunoprecipitation on selected novel damage-sensitive strains to better elucidate their roles in DNA damage repair. Here we demonstrate MMS sensitivity for ΔAzf1, ΔHtz1, ΔGat3, ΔLeo1, ΔIki3, ΔRic1, and ΔPop2, several yeast deletion mutants not previously reported to be sensitive to MMS. We show that, although the transcriptional profiles display broad similarities, the profile of ΔRad50 is unique. Finally, we propose roles for YML081W and Iki3 in DNA damage repair.

BACKGROUND

The integrity of genomic DNA is essential to the survival of cells and organisms. The DNA of cells is broken and rejoined during meiotic crossover, but cells must also repair DNA lesions caused by radiation or chemical insults. The basic cellular machinery that repairs DNA is highly conserved in eukaryotes from yeast to human. Defects in DNA damage repair pathways in humans are associated with high rates of cancer: notable examples include the BRCA1 mutation [164] and ataxia telangiectasia [165], which are defects in double-stranded break repair, and xeroderma pigmentosum [166], which is a defect in base excision repair.

Cells have two major DNA repair pathways: excision repair and double-stranded break repair. In excision repair, nucleotides that are damaged or mismatched are excised and replaced with correct, intact nucleotides. Double-stranded breaks can be repaired by non-homologous end joining (NHEJ), in which the broken ends are trimmed and re-ligated, or homologous recombination (HR), in which the homologous chromatid is used as a template for synthesis of a new strand. NHEJ may result in gain or loss of nucleotides due to end processing, or incorrect ligation may result in chromosome translocations. Thus, it is sometimes referred to as error-prone repair, whereas HR is also called error-free repair.

The primary causes of DNA damage are radiation and chemical damage [167]. Ultraviolet radiation damages DNA by causing pyrimidine dimers, which are repaired by photolyases or excision repair; ionizing radiation can damage DNA directly, by causing double-stranded breaks, or indirectly, by creating reactive oxygen species which then cause DNA lesions. In addition to reactive oxygen species, other chemicals that cause DNA lesions include alkylating and cross-linking agents. DNA lesions that result in mismatches are repaired by excision repair. Bulkier lesions, such as cross-links, may result in stalling or collapse of DNA replication forks, resulting in double-stranded breaks which are repaired by homologous recombination [167].

Double-stranded break repair

A double-stranded break is more complicated to repair than a base mismatch, and failure to repair it properly will likely result in chromosomal translocations, large deletions, or cell death. Cells contain groups of proteins to sense double-stranded breaks, arrest the cell cycle until damage is repaired, and repair the damage [167]. Many of these proteins are diffused throughout the nucleus, but when DNA damage is sensed, they form foci at the site of damage [168].

In yeast, double-stranded breaks are sensed by the MRX complex, which consists of the proteins Mre11, Rad50, and Xrs2. The MRX complex is the first set of proteins to arrive at damage foci [168, 169]. Mre11 exhibits 5' -> 3' exonuclease activity [170] and trims broken DNA ends [171]. Rad50, a member of the Structural Maintenance of Chromosomes (SMC) family [172], contains helicase domains and participates in DNA unwinding [173, 174]. Xrs2 helps direct the MRX complex to DNA ends and stimulates the exonuclease activity of Mre11 [175]. Mre11, Rad50, and Xrs2 all contain DNA-binding domains [171, 173, 175, 176]; both Mre11 and Rad50 are essential for the formation of DNA damage foci, but Xrs2 is not [168]. The MRX complex is necessary for both homologous recombination and non-homologous end joining. Homologous recombination, because it is more accurate, is the preferred method for double-stranded break repair. However, because Mre11 requires the kinase Cdk1 for proper end resection, and Cdk1 is inactive until the cell is committed to S phase [177], damage is repaired by NHEJ in the G1 phase of the cell cycle and by HR in the S and G2 phases [167].

In NHEJ, after the MRX complex senses damage, end processing of DNA is minimal. The Ku heterodimer (Ku70 and Ku80) binds the broken ends, and they are re-ligated by the DNA ligase IV complex, which includes the proteins Dnl4, Lif1 and Nej1 [178].

In homologous recombination, the MRX complex trims the DNA ends near the break via 5' to 3' end processing. Replication protein A, a heterotrimer of the proteins Rfa1, Rfa2, and Rfa3, binds to the 3' single-stranded DNA, protecting it from degradation and preventing formation of secondary structures. Accumulation of RPA causes recruitment of Mec1 and Ddc2 to the site of the lesion, as well as a sliding DNA clamp (Ddc1, Rad17, and Mec3) and the clamp loader (Rad24, Rfc2, Rfc3, Rfc4, and

Rfc5) [178]. Mec1, the clamp loader, and the clamp activate the DNA damage checkpoint [179]. Part of this checkpoint activation is accomplished when Mec1 and Tel1 hyperphosphorylate Rad9 [180, 181], which in turn causes Mec1 to phosphorylate Rad53 and Chk1 [182, 183]. This leads to cell cycle arrest until the damage is repaired.

Another role of Tel1 and Mec1 is to phosphorylate the histone H2A around double-stranded breaks. This results in recruiting of cohesins and chromatin remodelers such as Ino80 to the site of repair [184]; Mec1 also phosphorylates the Ies4 subunit of Ino80 [185]. These remodeling enzymes will open the chromatin to allow access by the homologous recombination machinery.

The actual homologous repair proteins are the last to arrive at the double-stranded break [168]. Rad51 replaces the RPA bound to the single-stranded DNA. Then Rad52, Rad55, and Rad57, which are the homologous recombination machinery, help Rad51 to find, open, and base pair with the homologous sequence. This base pairing is extended by DNA polymerase, and the nicks and gaps are then repaired by DNA ligase [167].

Experimental Strategy

One of the most useful methods to determine the function of a gene is to delete it and observe the organism. Essential genes are lethal when deleted, but the effects of non-lethal deletions may or may not be immediately apparent. Because numerous genes are only needed under specific circumstances, it may be necessary to subject the organism to a particular environmental perturbation for the role of the gene to become evident.

Many genes involved in homologous recombination in yeast are not lethal when knocked out and cause no obvious phenotype under normal growth conditions. They are often, however, sensitive to the alkylating agent methyl methane sulfonate (MMS). Therefore, to discover novel genes involved in DNA repair pathways and learn more about their functions, we screened a library of ~350 yeast deletion mutants known to bind

DNA or to be transcription factors for sensitivity to MMS. We chose 11 novel mutants for microarray transcriptional profiling and chromatin immunoprecipitation under MMS treatment.

Mutants

Adr1

Adr1 is a zinc finger transcription factor that regulates carbon source use; it is necessary for transcription of genes that allow yeast to grow in ethanol and glycerol [186, 187]. Adr1 has four transcription activation domains [188], and it binds to the consensus motif GG(A/G)G [189]. Its promoter contains binding sites for itself and for Cat8 [190], another transcription factor that regulates carbon source use [191]. The ΔAdr1 yeast deletion strain has previously been reported to be MMS-sensitive [192] and sensitive to the DNA cross-linking agent mitomycin C [193].

Asf1

Asf1 is member of the replication-coupling assembly factor (RCAF) complex, which re-assembles chromatin after double-stranded breaks have been repaired [194]. Under normal cell conditions, Asf1 forms a complex with Rad53, but when DNA is damaged or replication has stalled, it is released to promote nucleosome assembly by interacting with acetylated H3 and H4 [195]. The promoter of Asf1 contains binding sites for Skn7, Mbp1, and Swi6 [190]. The ΔAsf1 yeast deletion strain has previously been reported to be sensitive to MMS, hydroxyurea, the DNA topoisomerase I inhibitor camptothecin, and the DNA cross-linking agent cisplatin [196-199].

Azf1

Azf1 is a zinc-finger transcription factor which activates the G1 cyclin Cln3 when glucose is abundant; it binds to the motif AAGAAAAA [200]. In a two-hybrid screen,

Azf1 was found to interact with Rim1, a gene necessary for maintenance of the mitochondrial genome [201]. To the best of our knowledge, the ΔAzf1 yeast deletion strain has not been previously reported to be MMS-sensitive, but it has been reported sensitive to the cross-linking agents carboplatin and mitomycin C [193].

Cad1

Cad1 is a bZIP transcription factor that regulates the transcription of stabilizing and folding proteins during oxidative stress [202]. It binds to the consensus motif TTAGTAA [203]. The ΔCad1 yeast deletion strain has previously been reported as sensitive to MMS [192], cisplatin, and carboplatin [193].

Fkh2

Fkh2 belongs to the Forkhead transcription factor family. It cooperates with Mcm1 to activate Clb2, Swi5, and other Clb2 cluster genes necessary for the G2/M transition of the cell cycle [204]. Fkh2 binds to the consensus sequences (G/A)(T/C)(C/A)AA(C/T)A [205] and TGTTTNC [17]. The ΔFkh2 yeast deletion strain has previously been reported as sensitive to MMS [192], hydroxyurea, mitomycin C, and cisplatin [193].

Gat3

Gat3 is a 141-amino-acid protein with zinc finger domains from the GATA transcription factor family [206]. Deficiency in both the Hsp90 chaperone Hsp82 and in Gat3 produces a synthetic growth defect [207]. To the best of our knowledge, the ΔGat3 yeast deletion strain has not been previously reported MMS-sensitive, but it has been reported sensitive to mitomycin C [193].

Htz1

Htz1 codes for the variant histone H2A.Z, which is sometimes substituted for H2A by the SWR1 complex [208]. This substitution is more frequent in promoters than in coding regions; it correlates with reduced transcription, possibly by inhibiting the transcription-promoting histone modifiers Dot1, Set2, and NuA4 [192]. The ΔHtz1 yeast deletion strain has decreased resistance to hydroxyurea [209] and has been reported sensitive to mitomycin C [193] but, to the best of our knowledge, it has not been previously reported MMS-sensitive.

Iki3

Iki3 is a member of the Elongator complex, which is a part of the RNA polymerase II holoenzyme [210]. The Elongator complex acetylates the N-terminal tails of histones, primarily lysine #14 of H3 and lysine #8 of H4 [211]. To the best of our knowledge, the ΔIki3 yeast deletion strain has not been previously reported MMS-sensitive.

Leo1

Leo1 is a member of the Paf complex, which associates with RNA polymerase II during elongation. The Paf complex recruits and interacts with methyltransferases such as the COMPASS complex; thus, it is necessary for methylation of lysines #4 and #79 of histone H3 [212]. Leo1 is synthetically lethal with Vps72, the member of the SWR1 complex that binds Htz1, and with Yaf9, which is a subunit of both the SWR1 and NuA4 complexes [212]. To the best of our knowledge, the ΔLeo1 yeast deletion strain has not been previously reported as MMS-sensitive, but it has been reported sensitive to cisplatin [193].

Pop2

Pop2 has been reported as a possible transcription factor involved in glucose derepression [213] and a nuclease and a member of the RNase D family [214]. To the best of our knowledge, the ΔPop2 yeast deletion strain has not been previously reported as MMS-sensitive, but it has been reported sensitive to bleomycin, which causes double-stranded DNA breaks, and to hydroxyurea [193, 209, 215].

Rad50

Rad50 is a subunit of the MRX complex and a member of the Structural Maintenance of Chromosomes (SMC) family [172]; it contains helicase domains that unwind DNA [173, 174], and DNA-binding domains [173]. Rad50 is required for DNA damage focus formation [168]. ΔRad50 yeast deletion strains have been reported sensitive to MMS [216], camptothecin [216], hydroxyurea [199], and cisplatin [199].

Rad51

Rad51, a member of the Rad52 epistasis group, causes strand exchange between single and double-stranded DNA during homologous recombination [217]. The ΔRad51 yeast deletion strain has previously been reported sensitive to cisplatin [199], camptothecin [199], MMS [218], and hydroxyurea [215].

Rad52

Rad 52 acts as a Rad51 co-factor to promote efficient strand exchange during homologous recombination [219]. The ΔRad52 yeast deletion strain has previously been reported sensitive to MMS [218, 220], cisplatin [199], camptothecin [216], and hydroxyurea [199, 221].

Ric1

Ric1 appears to regulate the transcription of ribosomal proteins and rRNA [222]. It is also necessary for proper localization of trans-Golgi network proteins [223]. The ΔRic1 yeast deletion strain has previously been reported as sensitive to bleomycin [193] but, to the best of our knowledge, it has not previously been reported sensitive to MMS.

Rtt107

Rtt107 is phosphorylated by Mec1 when DNA is damaged. The absence of Rtt107 causes hypersensitivity to DNA damage during S-phase and inability to resume DNA replication after damage repair [224]. Rtt107 contains four BRCT domains [225], which are named for their similarity to the human BRCA1 gene and promote protein-protein interactions. The ΔRtt107 yeast deletion strain has previously been reported sensitive to hydroxyurea [196, 209, 215], campthotecin [193], and MMS [193, 218].

Sgs1

Sgs1 is an ATP-dependent DNA helicase [226] thought to be necessary for successful repair of certain homologous recombination intermediates [227]. It interacts with topoisomerase II and promotes proper chromosome segregation [228]. Sgs1 is necessary for genome stability; ΔSgs1 deletion strains exhibit mitotic hyper-recombination [229]. The ΔSgs1 yeast deletion strain has previously been reported as sensitive to MMS [218] and hydroxyurea [196, 209, 215].

Xrs2

Xrs2 is belongs to the MRX complex, but it is not essential for DNA damage focus formation [168]. It contains a DNA-binding domain, helps the MRX complex to find DNA ends, and stimulates Mre11 to trim them [175]. The ΔXrs2 yeast deletion

strain has been reported sensitive to cisplatin [199], camptothecin [199], and hydroxyurea [199, 209, 215] .

YML081W

YML081W is a gene of unknown function; the GFP fusion localizes to the nucleus [230, 231]. YML081W contains a zinc finger domain near the N-terminus. The protein sequence of YML081W is 41% similar [232] to Rsf2, a zinc-finger protein that regulates nuclear and mitochondrial genes, particularly those involved in glycerol-based growth and respiration [231]. Its zinc finger domain is similar to that of Adr1; this domain is also highly similar to the zinc finger domain of the Early Growth Response transcription factors (Egr1, Egr2, and Egr3) in organisms including human, mouse, chicken, zebrafish, and *Xenopus laevis* [232]. The remainder of the protein is similar to many other fungal proteins; with the exception of Mxrp1, which is necessary for methanol metabolism in *Pichia pastoris* [233], most of these have not been well-characterized (Supplemental Figure 4.1). Its promoter contains two binding sites for Abf1 and two binding sites for Reb1 [17]. Although both Rsf2 and Adr1 are required for growth on glycerol [187, 234], we did not observe any growth defect for ΔYML081W when plated on glycerol-based media. To the best of our knowledge, the ΔYML081W deletion strain has not previously been reported MMS-sensitive, although it has been found sensitive to mitomycin C [193].

MATERIALS AND METHODS

Yeast Defect Screening

A collection of ~350 DNA-binding yeast gene deletion strains from the Saccharomyces Genome Deletion Project (Open Biosystems) was screened for sensitivity to methyl methane sulfonate. Mutants were picked from freezer stock and grown in YPD in 96-well plates. Optical density at 600 nm was read with a plate reader. Cells were then resuspended in sorbitol (1 M) to a final OD of 0.1, and two tenfold serial dilutions were made. The three dilutions were spotted onto plates containing synthetic complete media + 0.02% methyl methane sulfonate. Plates were incubated for four days at 30°C and photographed daily. Yeast strains that were sensitive to MMS were streaked and spotted on YPD plates and on YPD + .02% MMS plates for phenotype verification (Figure 4.1). Gene deletion was verified by PCR as described in chapter 3. Eleven mutants whose roles in DNA damage repair were unclear and seven mutants with well-understood roles in damage repair were chosen for further study (Table 3.1).

Transcriptional Profiling

The parent strain BY4741 and the deletion strains ΔAdr1, ΔAsf1, ΔAzf1, ΔCad1, ΔFkh2, ΔGat3, ΔHtz1, ΔIki3, ΔLeo1, ΔPop2, ΔRad50, ΔRad51, ΔRad52, ΔRic1, ΔRtt107, ΔSgs1, ΔYML081W, and ΔXrs2 were grown, treated with MMS for one hour, and harvested as described in chapter 3. Culture volume was 50 mL. Total RNA isolation, reverse transcription, dye coupling, array hybridization, scanning, and normalization were as described in chapter 3. For each array, RNA from a deletion strain was used in the Cy5 channel and total RNA from BY4741 was used in the Cy3 channel. A minimum of two biological replicates was used for each strain. Custom Java software [158] was used to

combine replicates and calculate *P*-values for differential expression using the Error Model, as described in chapter 3. Differentially expressed genes between wild type and knockout strains were defined as those with $P < 0.001$.

Chromatin Immunoprecipitation

The TAP-tagged strains Adr1-TAP, Azf1-TAP, Iki3-TAP, Ric1-TAP, Rtt107-TAP, and YML081W-TAP were grown, treated with MMS for one hour, cross-linked with formaldehyde, and harvested as described in chapter 3. Culture volume was 200 mL. Chromatin immunoprecipitation, round A/B amplification, dye coupling, array hybridization, scanning, and normalization were as described in chapter 3. For each array, amplified ChIP DNA from the strain was used for the Cy5 channel and amplified genomic DNA from the strain was used for the Cy3 channel. A minimum of three biological replicates was used for each strain. Custom Java software [158] was used to combine replicates and calculate *P*-values for target genes more highly bound in the ChIP DNA than genomic DNA, using the error model as described in chapter 3. Target genes were defined as those with $P < 0.05$.

Clustering and Gene Ontology

Custom Java software [158] was used to determine enriched GO-Slim terms for deletion strains (Figure 4.3). Clustering for enriched GO terms was performed using the Saccharomyces Genome Database website [235]. Cluster [236] was used to group strains according to error model Z-score similarity (Figure 4.4) and average log2 ratio of replicates. Clusters were calculated for all spots and for higher-scoring subsets of spots. Z-score subsets consisted of those genes with Z scores of absolute value > 3, and genes with Z scores of absolute value > 4. In order to observe clustering based on genes that were very differentially expressed between the various mutant strains, two log2 ratio

subsets were defined as those in which the difference between the highest and lowest log2 ratios was > 2, and those in which the difference between highest and lowest log2 ratios was > 3.

Yeast Sequences and Annotation

Yeast annotation data and sequences of gene promoters were obtained from the Saccharomyces Genome Database [237] .

Motif Discovery

Perl scripts were used to extract upstream sequences (500 bp, 200 bp, and 100 bp) from genes differentially expressed in deletion strains, and from ChIP-chip target genes. Additional Perl scripts were used to run MEME [238], Alignace [239], and MD-Scan [240] motif discovery programs on these upstream sequences. EnoLOGOS [241] was used to visualize motifs.

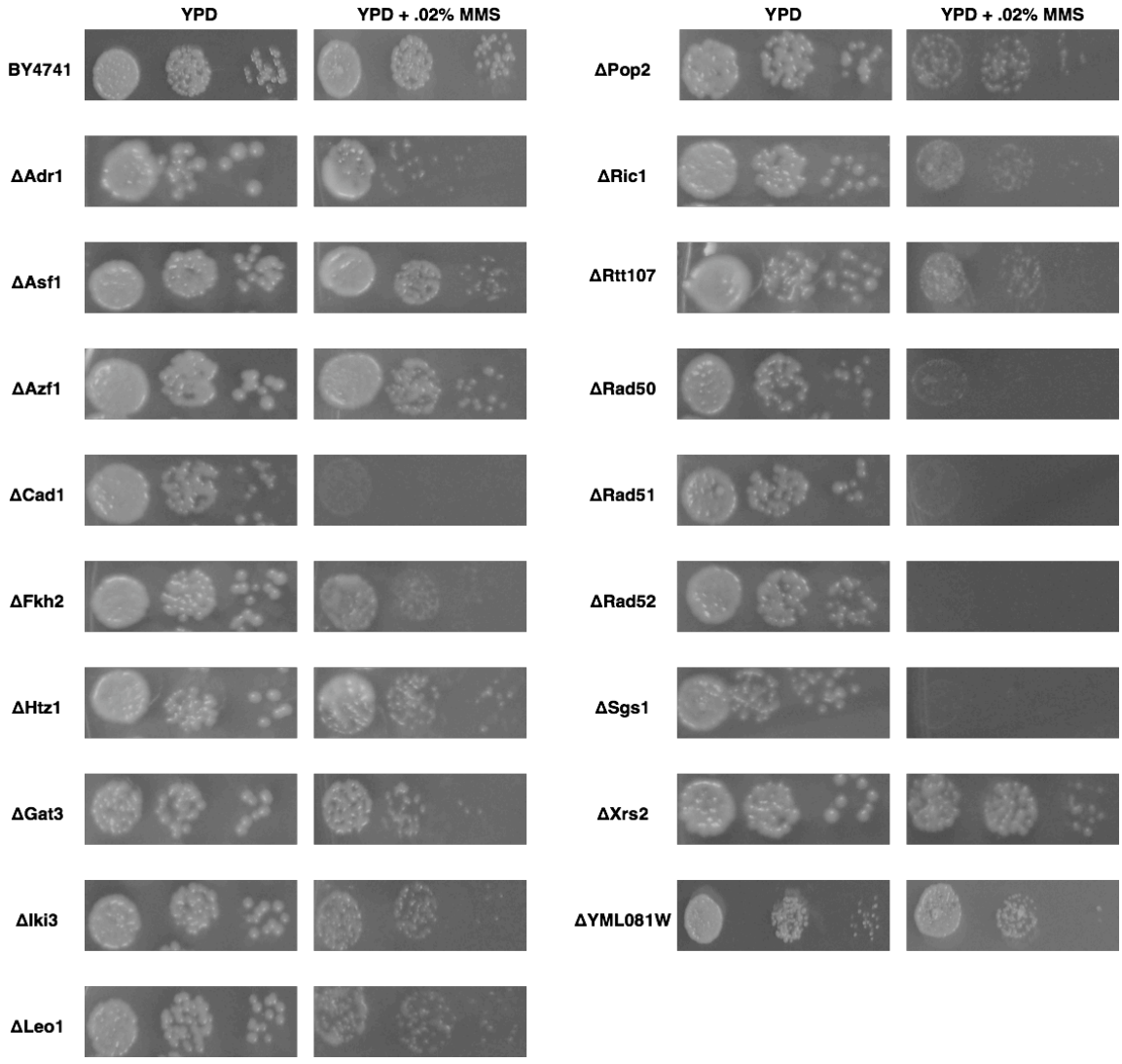


Figure 4.1: MMS-sensitive yeast deletion strains

Each deletion strain was grown overnight and diluted to OD 600 nm ~0.1. Tenfold serial dilutions were spotted onto YPD plates and YPD + 0.02% MMS plates. Plates were incubated at 30°C for 48 hours.

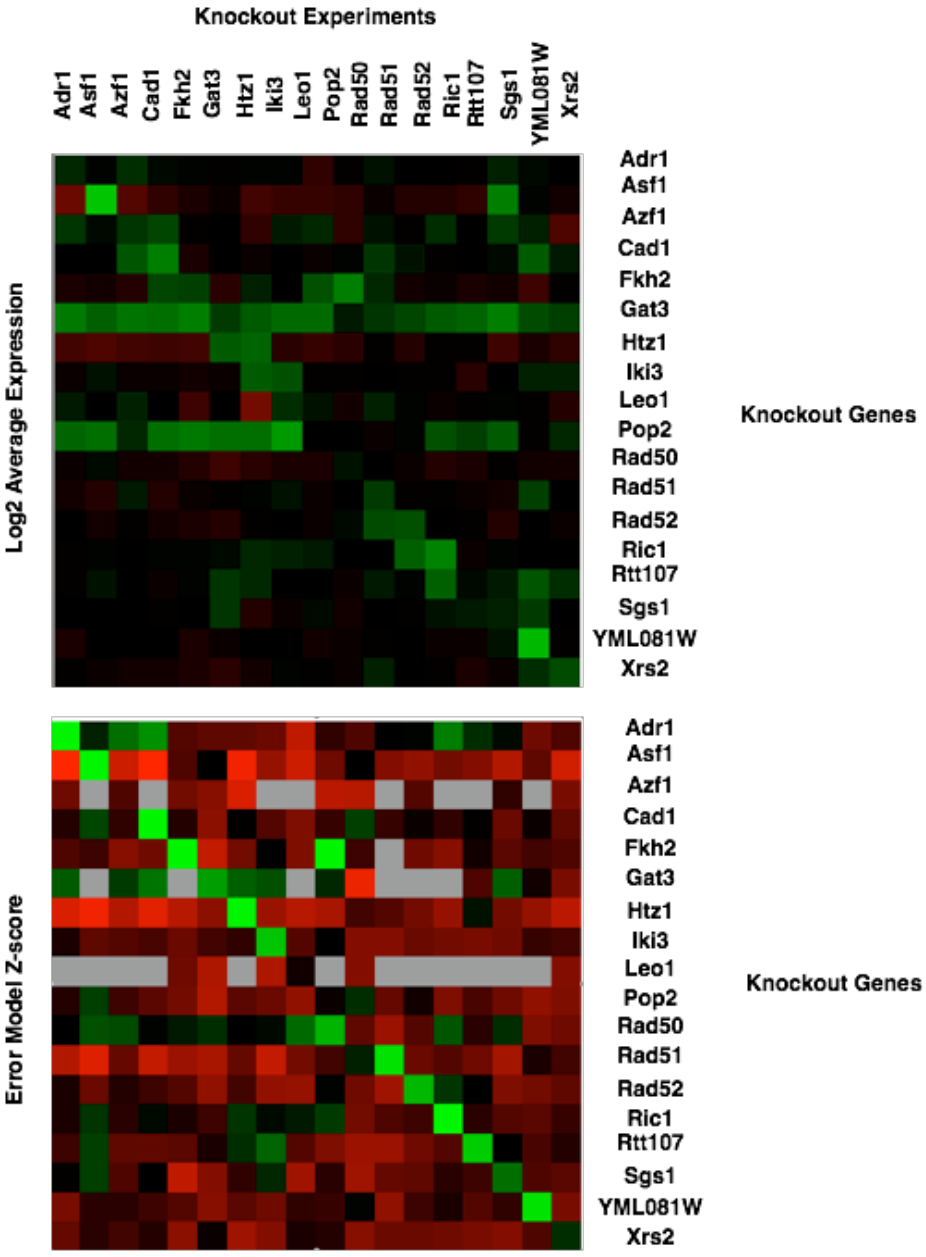


Figure 4.2: Quality Control

To assure quality control, biological replicate experiments were combined for each strain, and the vector of average log2 expression ratios (top) or error model-derived Z-score (bottom) was extracted for each gene. A diagonal line clearly indicates that deleted genes are not expressed and that these deletions are statistically significant.

Expression Experiment	**Hu KO**	**Workman KO**	**Workman KO + MMS**
Adr1	0.59	0.98	0.27
Azf1	0.22		
Cad1	0.84	0.99	**4.3x10^-5**
Fkh2	0.77	0.51	**1.7x10^-3**
Gat3	**0.02**		
Pop2	0.21		
Ric1	**1.09x10^-4**		
Rtt107	0.73		
YML081W	**0.02**		

ChIP experiment	Harbison YPD	Tachibana	Workman YPD	Workman MMS
Azf1-TAP	**0.02**			
YML081W-TAP	0.32			
Adr1-TAP	0.98	0.08	0.17	0.79

Table 4.1: Agreement of target sets with other data

We compared our expression target gene lists to those of Workman [242] and Hu [158], and our ChIP data targets to those of Harbison [17], Workman [242], and Tachibana [190]. Comparison data set targets were set at $P < 0.05$. Table 4.1 shows *P*-values of overlap significance as measured by the hypergeometric distribution. Overlap in differentially expressed genes was marginal in most experiments in which the strains were grown in YPD, but it was significant for strains treated with MMS.

RESULTS

Quality Control

All deletion strains were re-plated on YPD + .02% MMS plates to verify the sensitivity phenotype (Figure 4.1). To provide a visible assurance of gene knockout and expression profiling quality, we averaged log2 ratios of gene expression across experimental replicates for each deleted gene, then displayed them so that the list of deleted genes is on the Y axis, while the X axis contains averaged deletion strain experiments in which the particular gene is knocked out. Genes and experiments are aligned in alphabetical order on each axis, so the green diagonal shows that expression of each deleted gene is lower in the knockout strain than in the wild-type strain (Figure 4.2, top). To demonstrate the applicability of our error model statistical analysis, we created an identically-oriented display using Z-scores of error model-combined replicates rather than log2 ratios (Figure 4.2, bottom).

We compared our lists of differentially expressed genes in deletion experiments to those by Hu et al [158], in which deletion and wild-type strains were grown in YPD without MMS, and to those by Workman et al [242], which measured differential expression in a given strain after one hour of MMS treatment. We used a cutoff of $P < 0.05$ to select targets from both papers and calculated the hypergeometric probability of overlap between their sets of target genes and ours (Table 4.1, top). We found that overlap was significant in only three of nine strains from Hu. Although there was no significant overlap with between our targets and those from Workman's experiments without MMS, strikingly, overlap for two of the three strains became highly significant in experiments in which the strains were treated with MMS.

We further compared lists of ChIP-chip target genes to those of Harbison [17], Workman [242], and Tachibana [190] with $P < 0.05$ (Table 4.1, bottom). For all those experiments, yeast was grown in YPD media. The Workman dataset included ChIPs in media both with and without MMS. Agreement between our ΔAzf1 ChIP and the Harbison data was significant, but the ΔAdr1 and ΔYML081W data were not. The concurrence between the Harbison and Workman data is significant, but the Tachibana data agree significantly with neither set. Our ΔAdr1 data agrees more with the Tachibana dataset than with those of Workman and Harbison. As the Tachibana set correctly finds known Adr1 target Adh2 [243] where the others fail, it is likely the most accurate of the three.

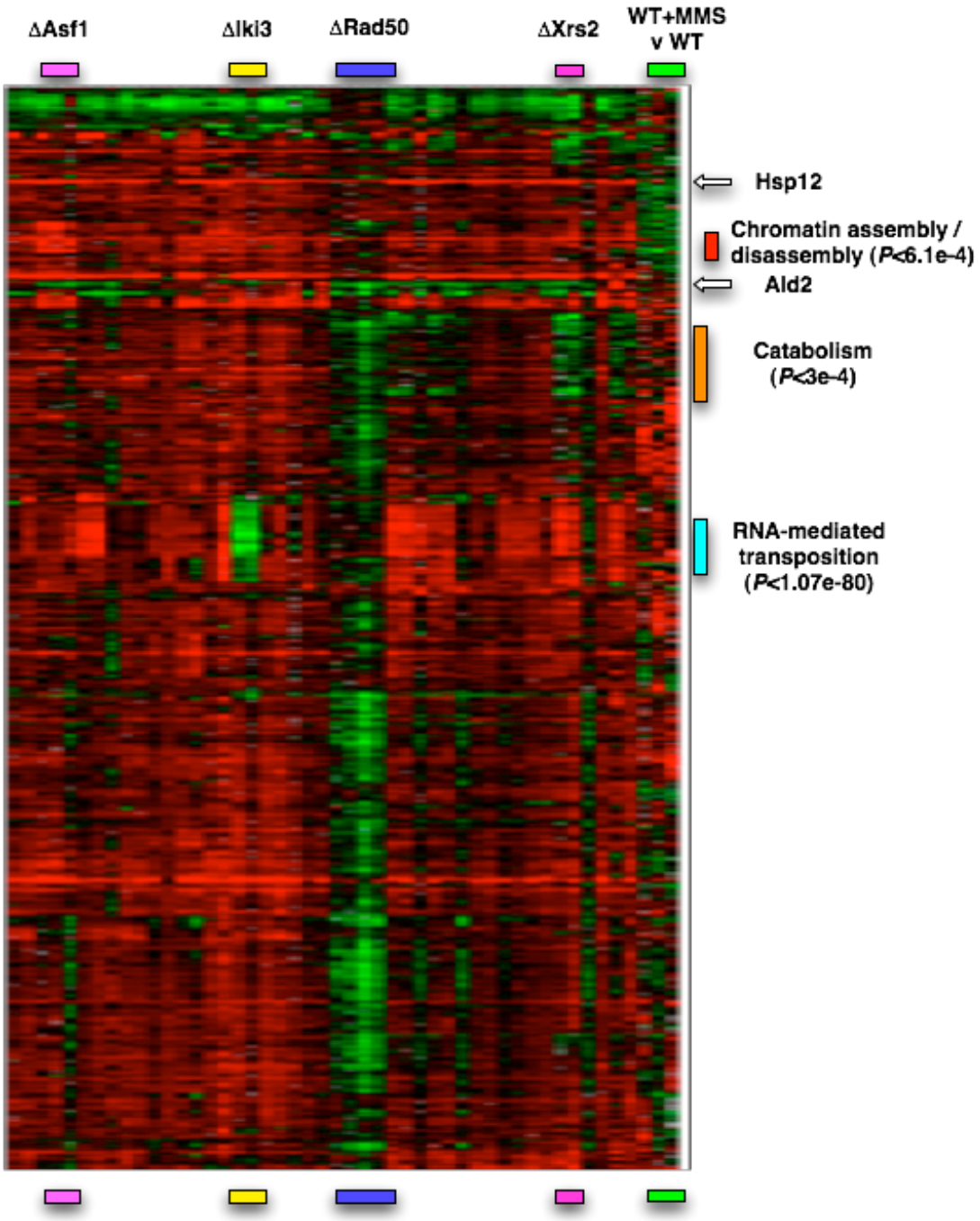


Figure 4.3: Expression patterns are broadly similar

The global pattern of gene expression is broadly similar between most mutant strains, but the expression profile of ΔRad50 is unique.

Gene Expression

Figure 4.3 shows the 1200 array spots with significant differential expression in at least one mutant, as measured by Z-score; supplementary figure 4.1 shows the 331 clustered array spots with a twofold change in expression in at least two experiments. The global pattern of gene expression was broadly similar across mutants. Some of the most consistently highly expressed genes were those related to and induced by stress; these genes were strongly and significantly induced across nearly all replicates (Supplemental Figure 4.2). These genes include the heat shock proteins Hsp12, Hsp31, and Hsp50, the DNA damage-responsive protein Ddr2, and the aldehyde dehydrogenases Ald2 and Ald3. The cluster of genes with the greatest decrease in expression, however, was not enriched for stress response.

Despite the broad similarity of transcriptional profiles, there were also noticeable differences between the strains. First, the profile of ΔRad50 was unique among strains, often behaving in a nearly opposite manner. This difference was consistent between biological replicates and microarray slide prints. Although the stress proteins Ddr2, Hsp12, Hsp31, Hsp150, Ald2, and Ald3 were among the most highly upregulated across all replicates, their upregulation was noticeably weaker in the Rad50 deletion strain (Supplemental Figure 4.2), suggesting that Rad50 is required for a full DNA damage-induced response.

Figure 4.3 shows a cluster enriched for catabolic genes. An increase in expression of catabolic genes is a part of the environmental stress response [244], as this process provides the cell with energy. The genes in this cluster have increased expression in most strains relative to wild type, but they are decreased in ΔRad50 and

ΔXrs2. This suggests that some genes involved in the DNA damage response require the activity of both Rad50 and Xrs2 for activation.

A cluster of genes related to chromatin assembly and disassembly was more highly expressed in the ΔAsf1 mutant than in the other mutants. Interestingly, this cluster included the core histones Hta1, Hta2, Hhf1, Hhf2, Hht1, and Hht2 (Figure 4.3, Supplemental Figure 4.3). Finally, the ΔIki3 mutant displayed significant downregulation of RNA-mediated retrotransposition in contrast to every other strain. This decrease is also visible in Supplemental Figure 4.2.

Gene Ontology Analysis

Analysis of enriched GO-Slim terms revealed broad patterns of cell stress. Genes related to stress response were upregulated in all eighteen strains; oxidoreductase-related genes were upregulated in seventeen of them. Genes related to energy production were broadly upregulated: carbohydrate metabolism and generation of precursor metabolites were enriched in almost all strains. Cellular respiration and mitochondrial genes were upregulated in just over half of strains, as were cell wall genes. Electron transport and protein catabolism were increased in a quarter of all strains.

Interestingly, genes related to metabolism were upregulated in half of strains. Amino acid metabolism was upregulated in two thirds of strains; upregulation of lipid metabolism was specific to the ΔGat3 strain, while upregulation of translation regulator activity and DNA binding was unique to the ΔCad1 strain. Peptidase activity was increased in the ΔIki3, ΔLeo1, and ΔCad1 strains. Both motor activity and microtubule-organizing genes are upregulated in ΔXrs2.

The deletion strains were clustered by GO-Slim profile similarity (Figure 4.4). The profiles of ΔXrs2, ΔRad51, ΔRad52, and ΔRtt107 are similar for both upregulated and downregulated genes, and the profile of ΔRad50 is the most unique. The clustering

partners for the remaining strains are inconsistent between upregulated and downregulated genes.

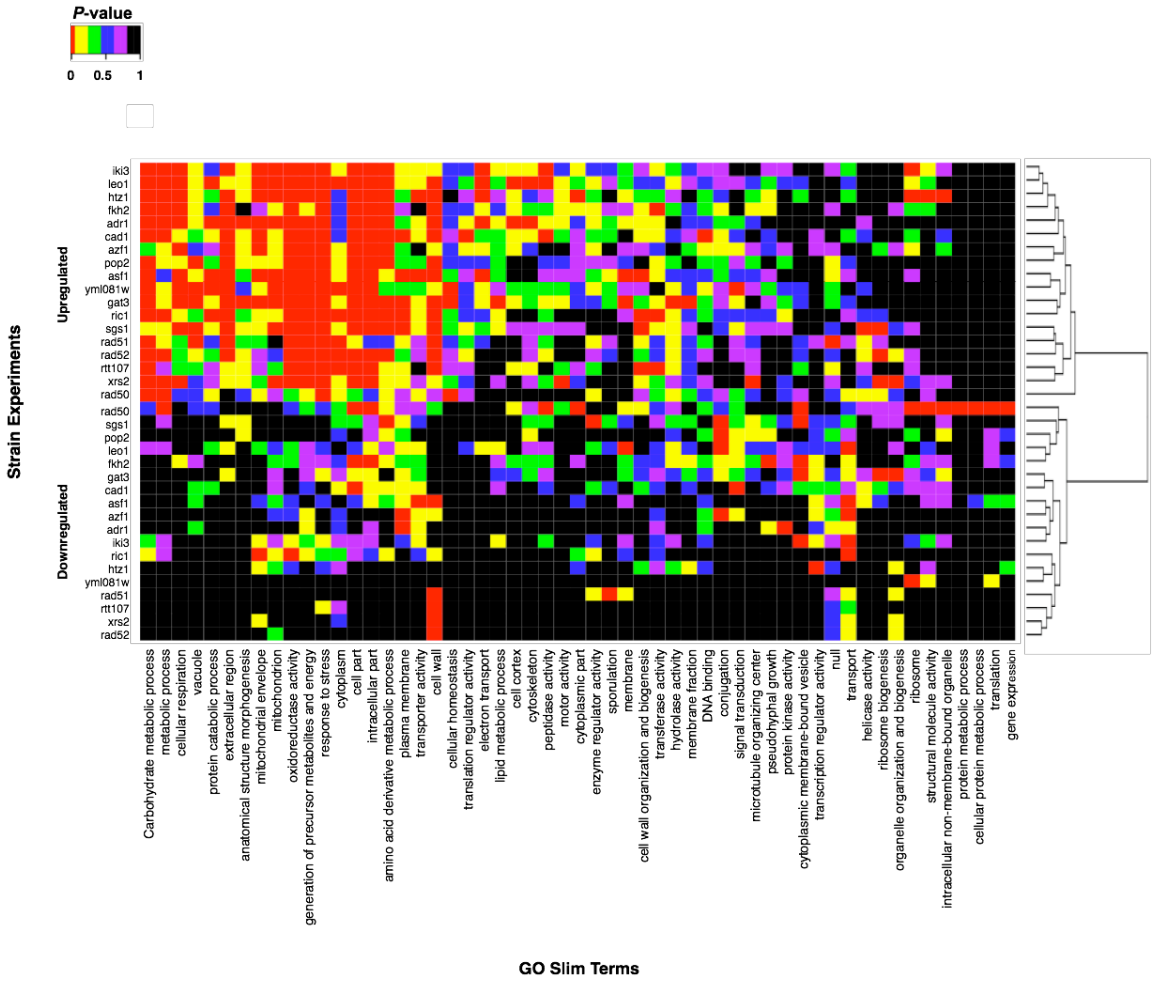


Figure 4.4: Enriched GO-Slim Terms

A heat map of enriched GO-slim terms among differentially expressed genes in each deletion strain. Values for which $P < 0.05$ are indicated in red. Deletion strains are clustered according to GO-slim profile similarity for upregulated (top) and downregulated (bottom) genes.

Clustering of deletion strains by expression profile

Whether strains were clustered by Z-score or by difference in log2 ratio, and whether all genes or only subsets were used, several clear trends were observed (Figure 4.5). First, the profiles of the ΔRad50 and wild-type strains were the most unique. This is consistent with earlier gene ontology analysis findings. Second, the profiles of ΔRad51, ΔRad52, and ΔXrs2 were similar to one another. When average linkage Z-score clustering was performed, or when complete linkage clustering was performed for all Z-scores, the profile of ΔRtt107 was similar to the profiles of ΔRad51, ΔRad52, and ΔXrs2. This is also consistent with gene ontology results.

The profiles of ΔCad1 and ΔAdr1 clustered together dependably, as did those of ΔFkh2 and ΔGat3. ΔYML081W, ΔLeo1, and ΔHtz1 consistently co-clustered in Z-score data, but this was less regular in log2 ratio clustering. ΔRic1 was often found with ΔSgs1, while ΔAzf1 was frequently found near ΔAsf1. Results for ΔIki3 and ΔPop2 were highly variable. Gene ontology results support the similarity of ΔAdr1 and ΔCad1 profiles as well as ΔLeo1/ΔHtz1 similarity.

Motif Analysis

Meme, AlignAce, and MDscan all recovered a motif similar to the canonical Adr1 motif GG(A/G)G from the upstream regions of genes differentially expressed in our ΔAdr1 experiments. Furthermore, this motif was enriched in the -500 bp upstream regions of these differentially expressed genes relative to upstream regions of all genomic features (Figure 4.6). We discovered an additional motif with MDscan that is similar to the stress response element, CCCCT [245, 246], suggesting possible co-regulation by Msn2 or Msn4.

Both Meme and AlignAce recovered a motif containing the consensus sequence CCNNGNGGNGNNC from upstream regions of genes differentially expressed in the ΔYML081W experiments (Figure 4.7). This consensus was significantly enriched in these upstream regions when compared to the upstream regions in the entire yeast genome, as measured by the hypergeometric distribution. Interestingly, AlignAce also found this consensus in the expression data for ΔAzf1, ΔGat3, ΔIki3, ΔRad51, and ΔRtt107, and it was substantially enriched in these strains and in ΔHtz1 (Table 4.2).

AlignAce did not find useful motifs from the ChIP data. The only motif MDscan found in the ChIP data was "GCGATGAG," which was modestly enriched in the Iki3 ChIP data, but not in the corresponding expression data. This is more likely a motif for an Iki3 cofactor than for Iki3 itself.

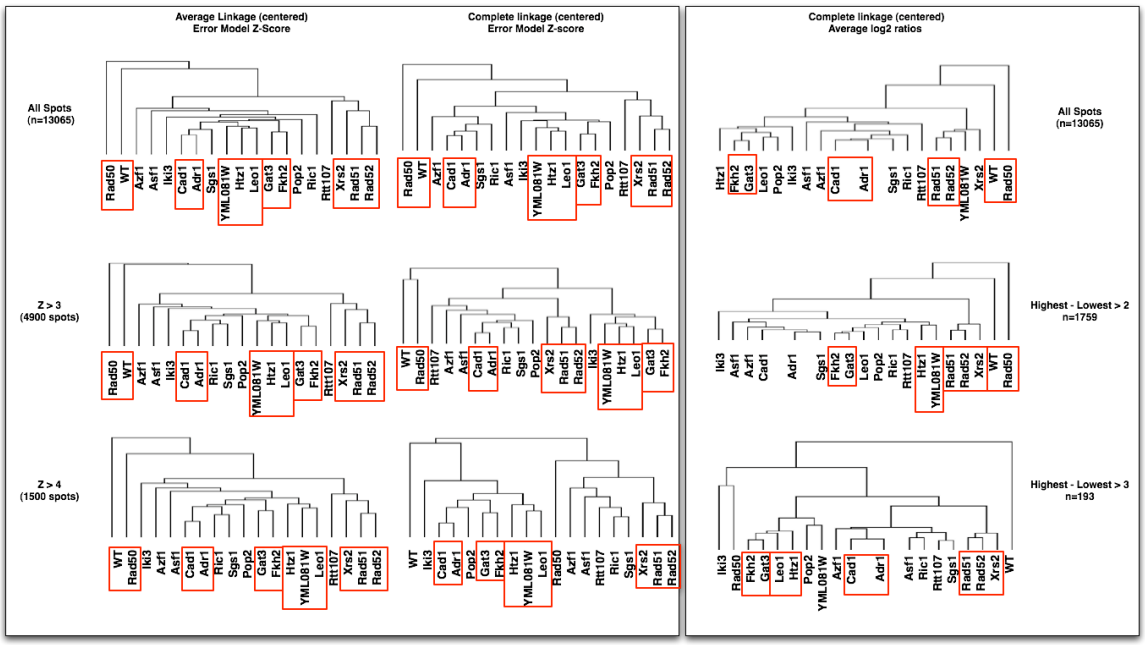


Figure 4.5: Co-Clustering

Deletion strains were clustered according to error model Z-score and according to average log2 ratio of replicates.

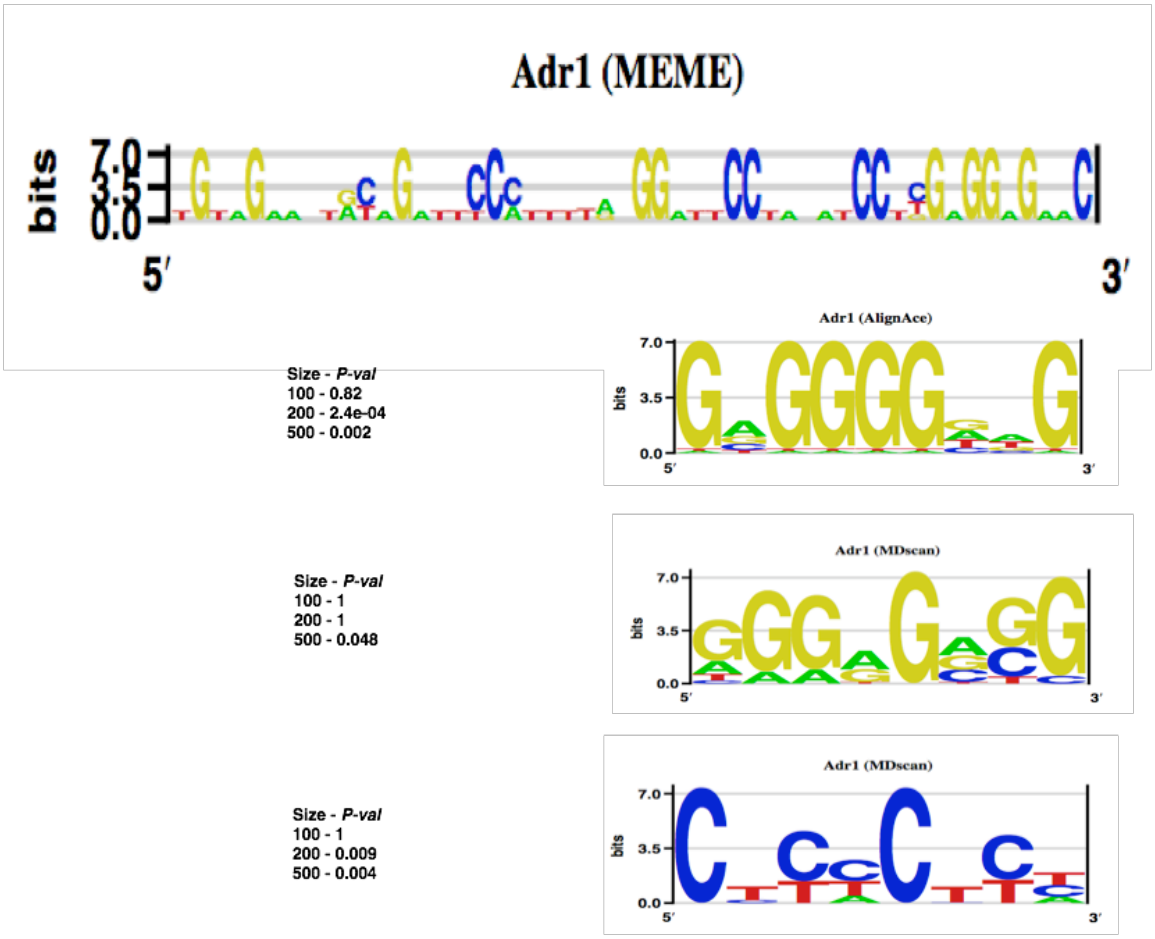


Figure 4.6: Results of Adr1 motif analysis

Meme, MDScan, and AlignAce discovered motifs similar to the Adr1 canonical motif, GG(A/G)G, in the upstream regions of genes differentially expressed in our ΔAdr1 expression data. These motifs were enriched in the upstream regions of ΔAdr1 target genes relative to their frequency in the upstream regions of all genomic features.

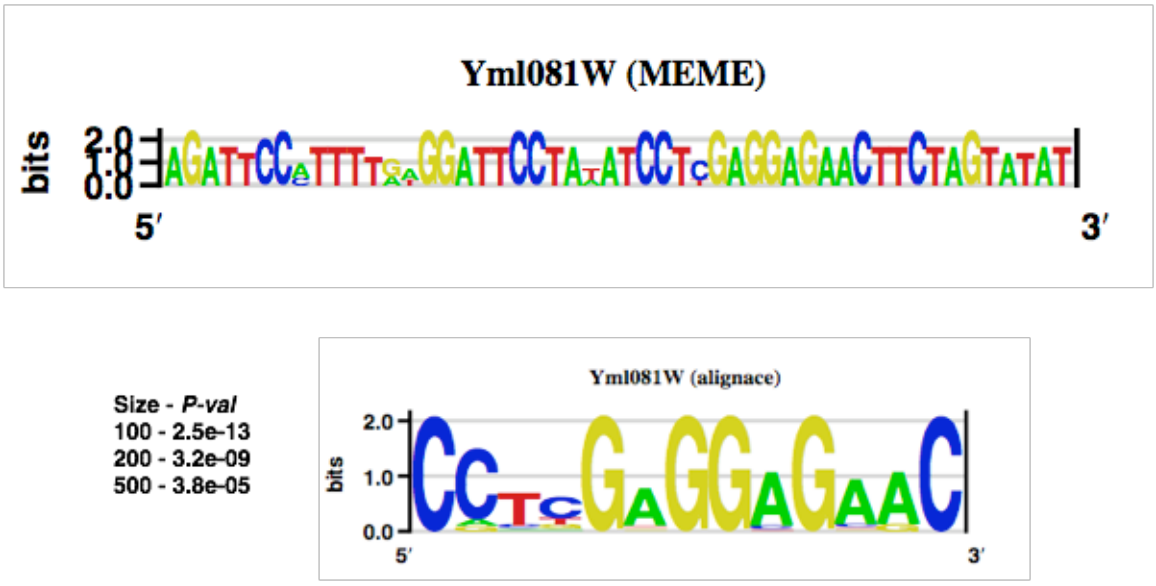


Figure 4.7: Results of YML081W motif analysis

Meme and AlignAce discovered a similar motif in the upstream regions of genes differentially expressed in our ΔYML081W expression data. A consensus of this motif (CCNNGNGGNGNNC) was enriched in the upstream regions of ΔYML081W target genes relative to their frequency in the upstream regions of all genomic features.

Promoter Size	TF	*P*-value	Promoter Size	TF	*P*-value
100	Adr1	1.959E-01	200	Leo1	8.456E-01
200	Adr1	2.787E-01	500	Leo1	6.671E-01
500	Adr1	6.413E-02	100	Rad50	2.860E-01
100	Asf1	**1.526E-03**	200	Rad50	2.189E-01
200	Asf1	**8.491E-03**	500	Rad50	6.732E-01
500	Asf1	1.580E-01	100	Rad51	**2.702E-24**
100	Azf1	**2.900E-08**	200	Rad51	**2.246E-18**
200	Azf1	**1.139E-06**	500	Rad51	**3.483E-12**
500	Azf1	**2.883E-05**	100	Rad52	7.496E-02
100	Cad1	9.683E-01	200	Rad52	2.477E-01
200	Cad1	9.542E-01	500	Rad52	2.075E-01
500	Cad1	7.299E-01	100	Ric1	7.813E-01
100	Fkh2	**2.831E-08**	200	Ric1	8.023E-01
200	Fkh2	**3.055E-06**	500	Ric1	8.323E-01
500	Fkh2	**2.955E-03**	100	Rtt107	**2.985E-23**
100	Gat3	**1.533E-03**	200	Rtt107	**3.182E-18**
200	Gat3	2.335E-02	500	Rtt107	**6.188E-11**
500	Gat3	1.822E-01	100	Sgs1	**5.680E-03**
100	Htz1	**6.223E-05**	200	Sgs1	7.885E-02
200	Htz1	**7.471E-04**	500	Sgs1	1.132E-01
500	Htz1	**1.435E-03**	100	Xrs2	6.993E-03
100	Iki3	**2.949E-11**	200	Xrs2	5.959E-02
200	Iki3	**2.265E-08**	500	Xrs2	1.345E-01
500	Iki3	**7.072E-05**	100	YML081w	**2.462E-13**
100	Leo1	5.396E-01	200	YML081w	**3.172E-09**
			500	YML081w	**3.778E-05**

Table 4.2: YML081W consensus enrichment in other target genes

The consensus CCNNGNGGNGNNC is enriched in differentially expressed genes corresponding to many of the mutants used in this study.

Genes Regulated by YML081W

The YML081W ChIP targets from our experiments overlapped the expression targets with P=0.011 as measured by the hypergeometric distribution. These differentially expressed genes and ChIP target genes included 9 of 98 immediate YML081W neighbors in YeastNet version 3.0 [247-249].

Among the ChIP targets of YML081w were numerous genes involved in DNA repair. These genes include Vps72 and Swc3, members of the SWR1 complex, which binds and exchanges histone H2AZ for histone H2A [250, 251]; Thi4 and Mmg101, which are necessary for maintenance of the mitochondrial genome [252, 253]; Rad5 and Mms2, which cooperate in post-replication DNA repair [254]; Pol3, the catalytic subunit of DNA polymerase δ, which is necessary for nucleotide excision repair and double-stranded break repair [255]; Rtt102, a component of the SWI/SNF and RSC chromatin remodeling complexes [256] with a role in chromosome segregation [257]; Swi4, which regulates transcription of G1-specific genes [258] such as those involved in DNA repair [259]; Rad55, which assists Rad51 in strand invasion during homologous repair [260]; and Srs2, a DNA helicase [261] that is part of the Rad6 DNA repair epistasis group [262].

Figure 4.8 shows, first, the 323 array spots corresponding to ChIP target genes and intergenic regions. These spots were collapsed into 206 ChIP target genes (A). Next, ChIP targets are shown with corresponding expression data (B), demonstrating that the ChIP targets are split approximately equally between genes upregulated and downregulated in the knockout. Finally, corresponding expression data is shown for genes related to DNA damage repair which were also ChIP targets (C). Mms2, Pol3, Srs2, Swc3, Vps72, Swi4, Thi4, and Rtt102 are clearly downregulated in the YML081W

deletion relative to BY4741 when both are treated with MMS. This is consistent with a model in which YML081W is a transcriptional activator of these genes. Of these damage genes, Rad55 and Vps72 contained the motif discovered in our motif analysis, GNGGNG, in their upstream regions, but Vps72 is the more strongly downregulated of the two, making it the most likely candidate for regulation by YML081W. Furthermore, chromatin remodelers are three of eight genes involved in DNA damage repair, bound by YML081W in ChIP, and downregulated when YML081W is deleted, and a role in chromatin remodeling is supported by the clustering of the ΔYML081W with ΔLeo1 and ΔHtz1. This does not, however, rule out a regulatory role for the other genes.

Figure 4.9 shows these candidate genes and their 50 nearest neighbors in YeastNet 2.0 [248]. Added connections suggested by the ChIP data are drawn in red.

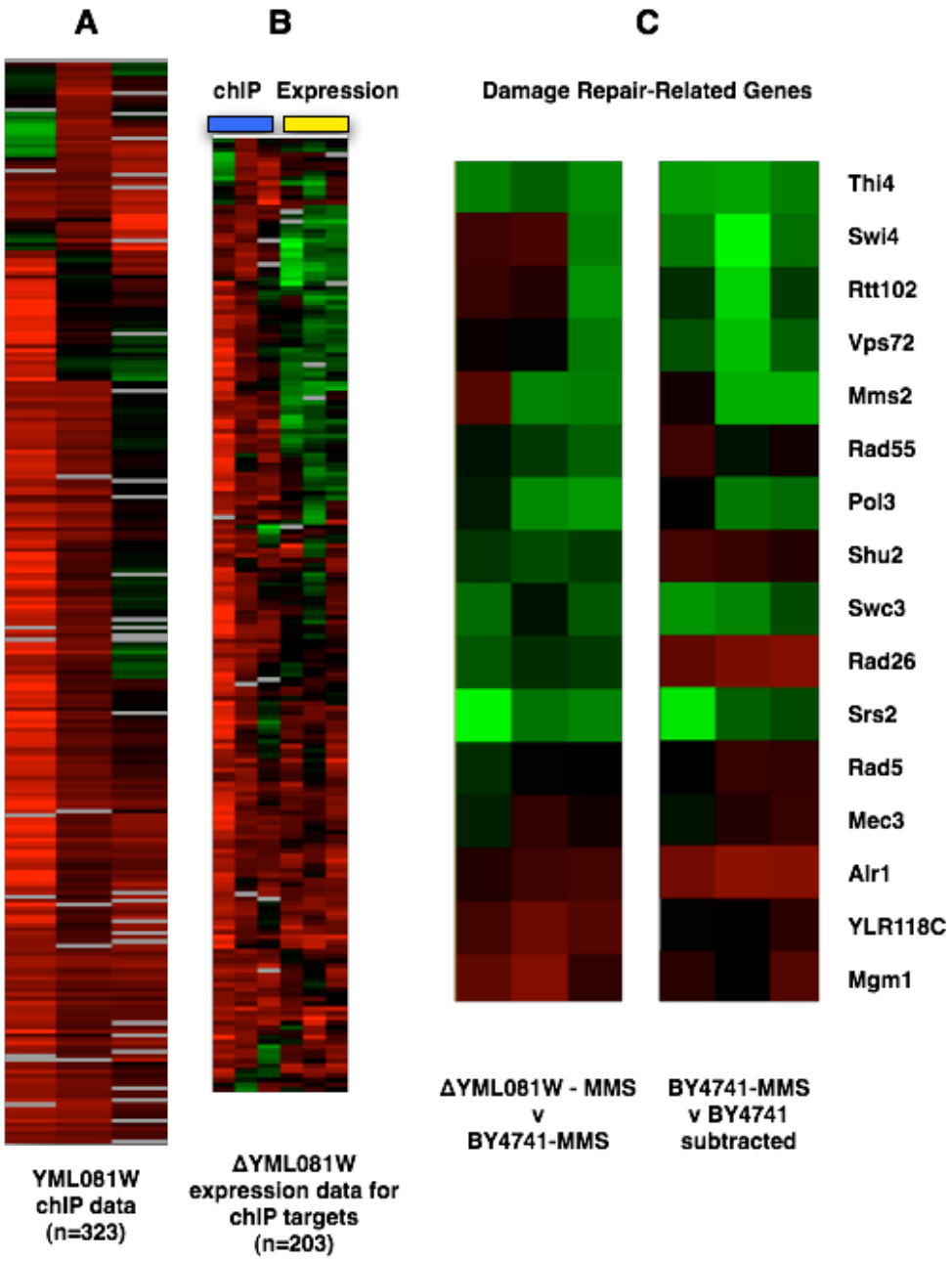


Figure 4.8: The YML081W-regulated damage set

A) ChIP target genes of YML081W, $P<0.05$. B) ChIP data (left) with corresponding expression data (right). C) Expression data for selected ChIP targets related to DNA damage repair. Data is shown as-is (left) and transformed by subtracting corresponding BY4741-MMS v BY4741 values. Genes with green are strong candidates for positive regulation by YML081W.

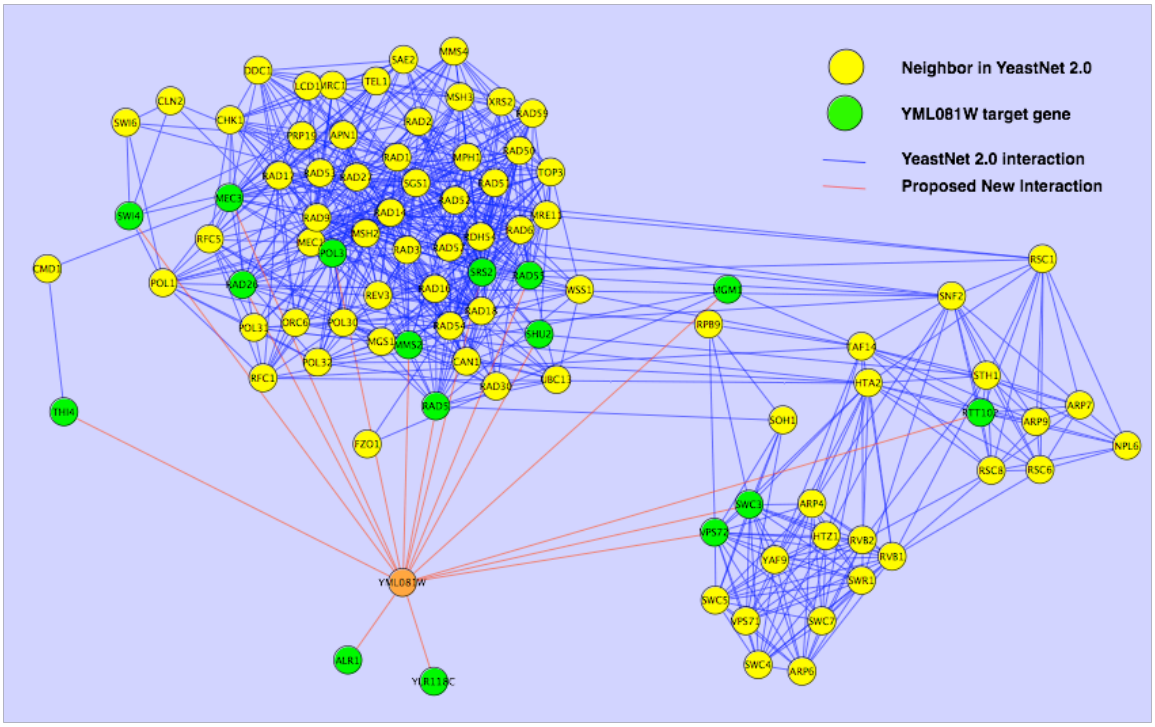


Figure 4.9: Proposed YML081W network

YML081W (orange) and its ChIP targets involved in damage repair (green) are shown with their 50 nearest neighbors in YeastNet 2.0 (yellow) [248].

The Role of Iki3

Iki3 is a DNA-interacting protein but lacks transcription factor domains such as zinc fingers. As such, it is most likely an indirect transcriptional regulator, and the motif found in the ChIP data likely belongs to a co-factor rather than to Iki3 itself. Iki3 had 184 ChIP targets at $P < 0.05$; they were roughly equally induced or repressed in the corresponding expression data (Figure 4.10A). Three ChIP targets, Rad34, Srs2, and Mum1, had functions related to DNA damage, but none contained the GCGATGAG motif in its promoter. However, Iki3 and Srs2 were found to interact in a two-hybrid screen [263], and Srs2 was strongly downregulated in our corresponding expression data, suggesting that Iki3 may be an indirect regulator of Srs2, and the DNA damage sensitivity of ΔIki3 may be caused by buildup of recombination intermediates due to lack of Srs2 helicase activity.

Ty retrotransposon genes were strongly and significantly repressed in ΔIki3, but in none of the other mutants (Figure 4.2, Figure 4.10C, Supplemental Figure 4.3). The consensus motif CCNNGNGGNGNNC, found in motif analysis of ΔYML081W target genes, is heavily overrepresented in differentially expressed genes in the ΔIki3 mutant (Table 4.2); this motif occurs almost exclusively in upstream regions of Ty retrotransposons. Previous work in our lab [158] found no significant difference in Ty gene expression between ΔIki3 and BY4741 when both were grown in YPD. As DNA damage is known to increase retrotransposition [264], these results suggest that ΔIki3 is necessary for Ty retrotransposon activity under DNA damage conditions.

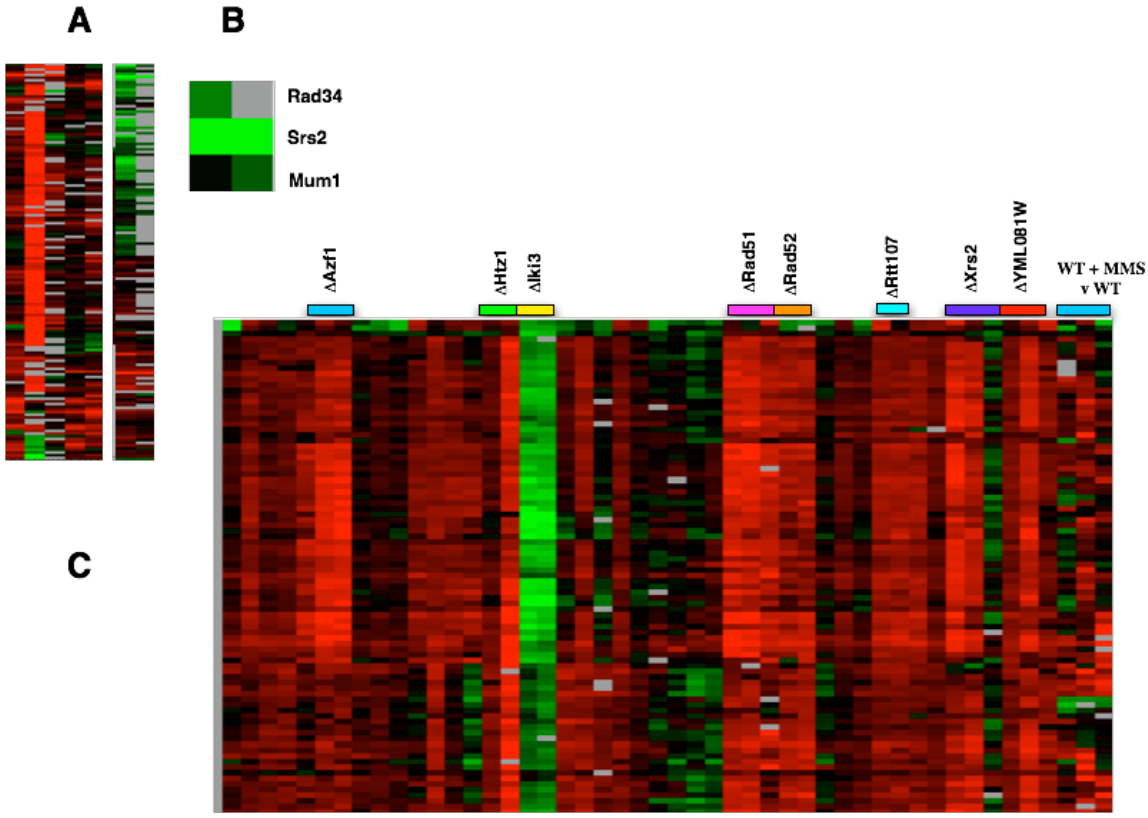


Figure 4.10: Iki3 expression and ChIP targets

A) Iki3 ChIP data (left) and corresponding expression data (right). B) Genes involved in DNA damage repair that were Iki3 ChIP targets and were downregulated in expression data. C) Retrotransposons are strongly and uniquely repressed in the ΔIki3 strain.

DISCUSSION AND CONCLUSIONS

Here we have described the screen of a library of yeast deletion strains for novel MMS-sensitive mutants. We have used transcriptional profiling, chromatin immunoprecipitation, and bioinformatics tools to predict their roles in DNA damage repair.

Novel Sensitive Mutants

As a result of our screen of 350 yeast deletion mutants for sensitivity to MMS, we have demonstrated MMS sensitivity for the yeast mutants ΔAzf1, ΔHtz1, ΔGat3, ΔLeo1, ΔIki3, ΔRic1, ΔPop2, and ΔYML081W, which were not previously reported as MMS-sensitive. With the exception of ΔIki3, all these mutants have previously been described as sensitive to other DNA-damaging chemicals such as hydroxyurea, bleomycin, cisplatin, or mitomycin C.

Expression Profiling

A well-known previous study [244] found that the damage-deficient ΔMec1 and ΔDun1 deletion strains had very similar transcriptional profiles when treated with MMS. This study found, furthermore, that activation of the Mec1 pathway was necessary for DNA damage to induce the environmental stress response (ESR), that induction of the stress transcription factor Msn4 was decreased in mutants, and that many genes that are typically repressed in the ESR were slightly induced in ΔMec1 mutants. For instance, the ribosomal genes might be slightly activated rather than repressed as expected [244].

Our data fit well with these findings. First, the expression profiles of most mutants are broadly similar. Activation of Mec1 kinase depends on the MRX complex [265]; furthermore, both Rad50 and Mre11 are required for the MRX complex to form

damage foci at DNA breaks [168]. Therefore, much of the large expression profile difference between the ΔRad50 strain and the other mutants may be due to failure of Mec1 activation and thus failure to induce the ESR by DNA damage. In supplemental figure 4.2, many stress response genes are strongly upregulated. In particular, the stress proteins Ddr2, Hsp12, Hsp31, Hsp150, Ald2, and Ald3 are among the most highly upregulated across all replicates, but this induction is noticeably lessened in the Rad50 deletion strain, suggesting, again, that Rad50 is required for a full damage response.

Our experimental setup, a direct comparison of mRNA levels of mutants versus wild type treated with MMS for one hour, has some distinct advantages. First, it allows for the use of a common reference. Second, it allows for a direct comparison to verify that deleted genes are deleted, as shown in Figure 4.2. Third, because both cells are undergoing the same stress, the gene expression differences should only be due to strain differences.

However, data must be interpreted carefully with this experimental setup because, as previously observed [244], the transcriptional response of a deletion mutant may be similar to a wild-type but to a lesser degree, or it may behave completely differently. For example, it has previously been found that histone transcription is reduced when wild-type yeast is treated with MMS [244], but we observe an increase in histone transcript levels in most of our mutants (Supplementary Figure 4.3).

The far-right columns of Supplementary Figure 4.3 (BY4741 + MMS vs BY4741) show that histone transcripts are downregulated when BY4741 is treated with MMS for one hour. RNA from ΔRad50 treated with MMS is hybridized to RNA from BY4741 treated with MMS, and histone transcripts are further modestly downregulated relative to wild type. However, when the wild-type strain shows downregulation and the mutants show upregulation, it cannot be ascertained whether the gene is actually

upregulated, or whether it is downregulated, but less strongly than in wild type. Data interpretation is therefore facilitated by availability of WT-MMS v WT data, but still challenging.

Gene Ontology and Co-Clustering

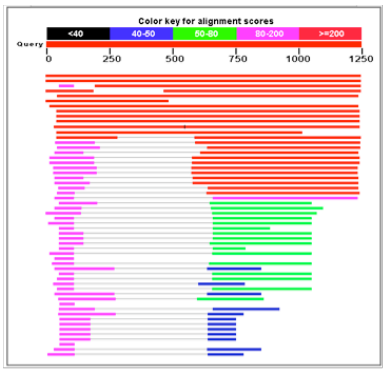
Our gene ontology analysis revealed a general stress response in the mutants that manifested in slightly different ways from strain to strain. The ΔRad50 strain was enriched in the smallest number of stress-related categories, which is consistent with ΔRad50 being necessary for full DNA damage-induced stress response. Clustering methods consistently classified ΔRad50 as the most dissimilar of the strains, correctly placed ΔRad51 and ΔRad52, members of the same epistasis group, together, and placed ΔXrs2 with them. Clustering strongly suggested greatest similarity between the ΔCad1 and ΔAdr1 mutants, the ΔGat3 and ΔFkh2 mutants, and between ΔHtz1, ΔLeo1, and ΔYML081W. Followup experiments might include ChIP of Gat3 to determine whether it binds the Cln2 cluster, or using fluorescence cell sorting to assay ΔGat3 for cell cycle defects.

YML081W and Iki3 Damage Roles

ChIP analysis revealed ΔYML081W binding to Vps72 and Swc3, members of the SWR1 complex, which exchanges H2AZ for H2A, and expression of both genes is reduced in ΔYML081W strains. Both YML081W and Vps72 are transcriptionally induced during nucleosome depletion [266]. Leo1 is a member of the Paf complex, which methylates histones, and it is synthetically lethal when combined with Vps72 deletion [212]. The co-clustering of ΔHtz1, ΔLeo1, and ΔYML081W therefore seems highly plausible (Supplemental Figure 4.4). In addition, the promoter of Vps72 contains the candidate motif we discovered for YML081W. These data all support a DNA

damage repair role for YML081W as a regulator of chromatin remodelers, although they do not rule out a role as a regulator of other YML081W ChIP target genes involved in repair of DNA damage, such as Mms2, Pol3, Swi4, Rtt102, and Thi4. Potential followup experiments might include immunoprecipitation of YML081W followed by mass spectroscopy or Western blot to look for co-immunoprecipitation with members of the SWR1 complex, or Western blots to measure differential protein abundance of SWR1 complex members in ΔYML081W and wild-type cells. These experiments could also be performed for Mms2, Pol3, Swi4, Rtt102, and Thi4. Pending the outcome of followup experiments, we propose to name YML081W "VPR1," for "VPS Regulator 1".

The most likely cause of ΔIki3 DNA damage defect appears to be via regulation of Srs2. This could be tested by attempting rescue of the ΔIki3 mutant by overexpression of Srs2.

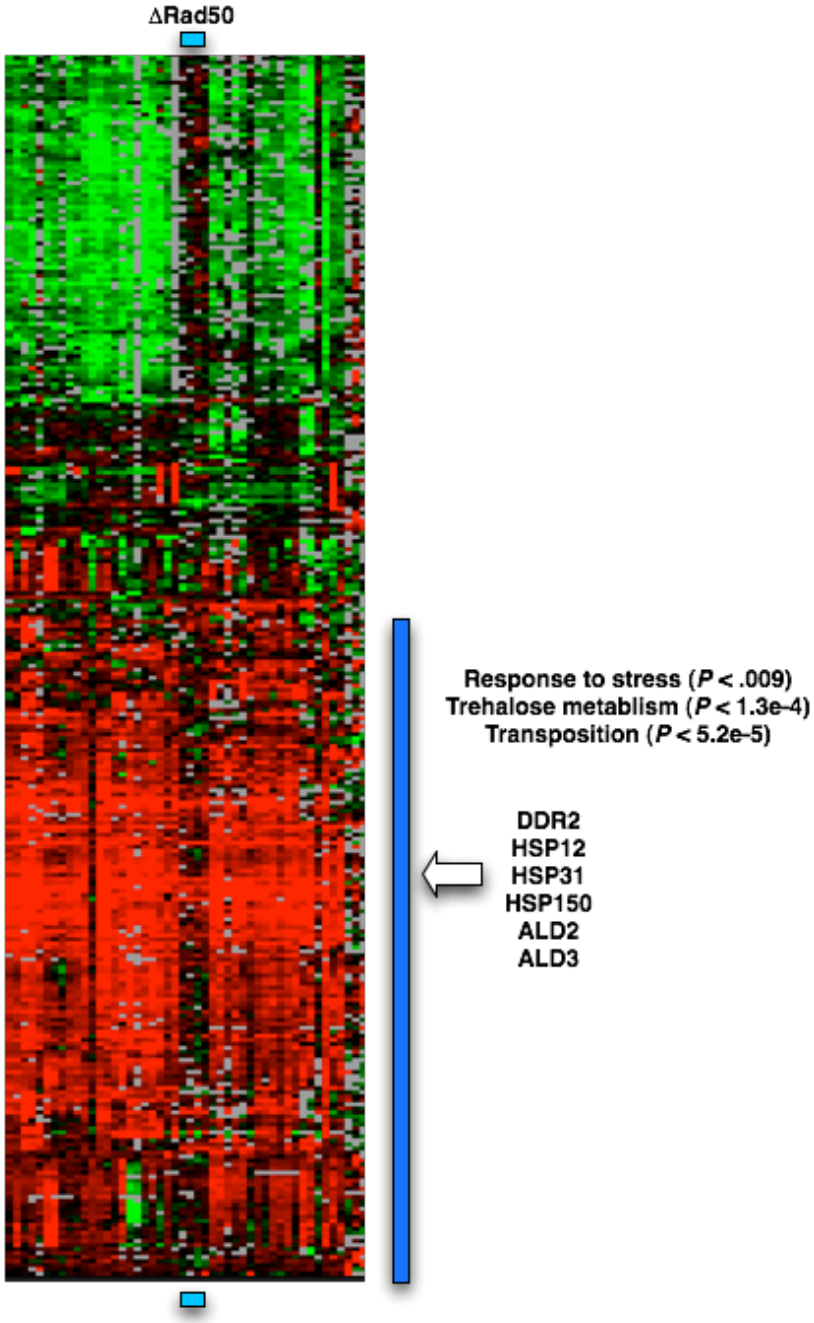


root
. Fungi/Metazoa group [eukaryotes]
. . Dikarya [fungi]
. . . Ascomycota [ascomycetes]
. . . . Saccharomycetales [ascomycetes]
. . . . . Saccharomycetaceae [ascomycetes]
. . . . . Saccharomyces cerevisiae [ascomycetes]

. . . . . Saccharomyces cerevisiae (yeast) ------	2581	37 hits [ascomycetes]	Putative protein of unknown function; green fluorescent pro
. . . . . Saccharomyces cerevisiae YJM789	2567	8 hits [ascomycetes]	conserved protein [Saccharomyces cerevisiae YJM789] >gi|190
. . . . . Saccharomyces cerevisiae RM11-1a	2567	8 hits [ascomycetes]	conserved protein [Saccharomyces cerevisiae YJM789] >gi|190
. . . . . Saccharomyces cerevisiae AWRI1631	2162	8 hits [ascomycetes]	YML081Wp-like protein [Saccharomyces cerevisiae AWRI1631]
. . . . . Ashbya gossypii ATCC 10895 ----------	972	4 hits [ascomycetes]	AER159Cp [Ashbya gossypii ATCC 10895] >gi|44983805|gb|AAS52
. . . . . Kluyveromyces lactis NRRL Y-1140	758	6 hits [ascomycetes]	unnamed protein product [Kluyveromyces lactis] >gi|49640864
. . . . . Kluyveromyces lactis	758	8 hits [ascomycetes]	unnamed protein product [Kluyveromyces lactis] >gi|49640864
. . . . . Vanderwaltozyma polyspora DSM 70294	615	16 hits [ascomycetes]	hypothetical protein Kpol_1001p1 [Vanderwaltozyma polyspora
. . . . . Nakaseomyces delphensis	446	2 hits [ascomycetes]	zinc-finger protein [Nakaseomyces delphensis]
. . . . . Debaryomyces hansenii	334	6 hits [ascomycetes]	DEHA2D06490p [Debaryomyces hansenii]
. . . . . Debaryomyces hansenii CBS767	331	6 hits [ascomycetes]	hypothetical protein DEHA0D07150g [Debaryomyces hansenii CB
. . . . . Pichia guilliermondii ATCC 6260	325	10 hits [ascomycetes]	hypothetical protein PGUG_02473 [Pichia guilliermondii ATCC
. . . . . Pichia stipitis CBS 6054	296	14 hits [ascomycetes]	Zinc Finger Protein C2H2-like protein [Pichia stipitis CBS
. . . . . Lodderomyces elongisporus NRRL YB-4239 ..	223	8 hits [ascomycetes]	conserved hypothetical protein [Lodderomyces elongisporus N
. . . . . Pichia pastoris	102	2 hits [ascomycetes]	methanol expression regulator I [Pichia pastoris]
. . . . . Pichia jadinii (Torula yeast)	70	1 hit [ascomycetes]	MIG repressor [Pichia jadinii]
. . . . . Candida glabrata CBS 138 ----------	1028	6 hits [ascomycetes]	unnamed protein product [Candida glabrata] >gi|49526287|emb
. . . . . Candida glabrata	1028	6 hits [ascomycetes]	unnamed protein product [Candida glabrata] >gi|49526287|emb
. . . . . Candida albicans SC5314	248	24 hits [ascomycetes]	potential zinc finger protein [Candida albicans SC5314] >gi
. . . . . Yarrowia lipolytica CLIB122	140	4 hits [ascomycetes]	hypothetical protein [Yarrowia lipolytica] >gi|49648865|emb
. . . . . Yarrowia lipolytica	140	4 hits [ascomycetes]	hypothetical protein [Yarrowia lipolytica] >gi|49648865|emb
. . . . Aspergillus clavatus NRRL 1 ----------	108	20 hits [ascomycetes]	C2H2 transcription factor (AmdX), putative [Aspergillus cla
. . . . Coccidioides immitis RS	107	8 hits [ascomycetes]	hypothetical protein CIMG_09396 [Coccidioides immitis RS] >
. . . . Neurospora crassa OR74A	105	8 hits [ascomycetes]	hypothetical protein NCU09496 [Neurospora crassa OR74A] >gi
. . . . Gibberella zeae PH-1	105	9 hits [ascomycetes]	hypothetical protein FG02743.1 [Gibberella zeae PH-1]
. . . . Chaetomium globosum CBS 148.51	105	6 hits [ascomycetes]	hypothetical protein CHGG_05487 [Chaetomium globosum CBS 14
. . . . Podospora anserina DSM 980	105	3 hits [ascomycetes]	unnamed protein product [Podospora anserina] >gi|170940418|
. . . . Podospora anserina	105	3 hits [ascomycetes]	unnamed protein product [Podospora anserina] >gi|170940418|
. . . . Phaeosphaeria nodorum SN15	104	4 hits [ascomycetes]	hypothetical protein SNOG_10626 [Phaeosphaeria nodorum SN15
. . . . Aspergillus fumigatus Af293	104	18 hits [ascomycetes]	C2H2 transcription factor (AmdX) [Aspergillus fumigatus Af2
. . . . Aspergillus fumigatus A1163	104	9 hits [ascomycetes]	C2H2 transcription factor (AmdX), putative [Aspergillus fum
. . . . Neosartorya fischeri NRRL 181	103	18 hits [ascomycetes]	C2H2 transcription factor (AmdX), putative [Neosartorya fis
. . . . Magnaporthe grisea 70-15	103	14 hits [ascomycetes]	hypothetical protein MGG_08757 [Magnaporthe grisea 70-15] >
. . . . Pyrenophora tritici-repentis Pt-1C-BFP	103	12 hits [ascomycetes]	conserved hypothetical protein [Pyrenophora tritici-repenti
. . . . Ajellomyces capsulatus NAm1	102	4 hits [ascomycetes]	hypothetical protein HCAG_02457 [Ajellomyces capsulatus NAm
. . . . Aspergillus terreus NIH2624	102	10 hits [ascomycetes]	hypothetical protein ATEG_10025 [Aspergillus terreus NIH262
. . . . Aspergillus niger CBS 513.88	102	8 hits [ascomycetes]	hypothetical protein An04g09290 [Aspergillus niger] >gi|134
. . . . Aspergillus niger	102	10 hits [ascomycetes]	hypothetical protein An04g09290 [Aspergillus niger] >gi|134
. . . . Aspergillus nidulans FGSC A4	102	12 hits [ascomycetes]	hypothetical protein AN9492.2 [Aspergillus nidulans FGSC A4
. . . . Emericella nidulans	102	5 hits [ascomycetes]	DNA binding regulatory protein AmdX [Emericella nidulans]
. . . . Botryotinia fuckeliana B05.10	98	10 hits [ascomycetes]	hypothetical protein BC1G_08399 [Botryotinia fuckeliana B05
. . . . Sclerotinia sclerotiorum 1980	95	10 hits [ascomycetes]	hypothetical protein SS1G_03567 [Sclerotinia sclerotiorum 1
. . . . Schizosaccharomyces pombe	88	12 hits [ascomycetes]	transcription factor Rst2 [Schizosaccharomyces pombe] >gi|2
. . . . Aspergillus oryzae RIB40	71	2 hits [ascomycetes]	hypothetical protein [Aspergillus oryzae RIB40] >gi|8377119
. . . . Aspergillus oryzae	71	2 hits [ascomycetes]	hypothetical protein [Aspergillus oryzae RIB40] >gi|8377119
. . . . Talaromyces emersonii	71	1 hit [ascomycetes]	catabolite repressor CreT [Talaromyces emersonii]
. . . . Aspergillus aculeatus	71	2 hits [ascomycetes]	DNA-binding protein creA (Carbon catabolite repressor) >gi|

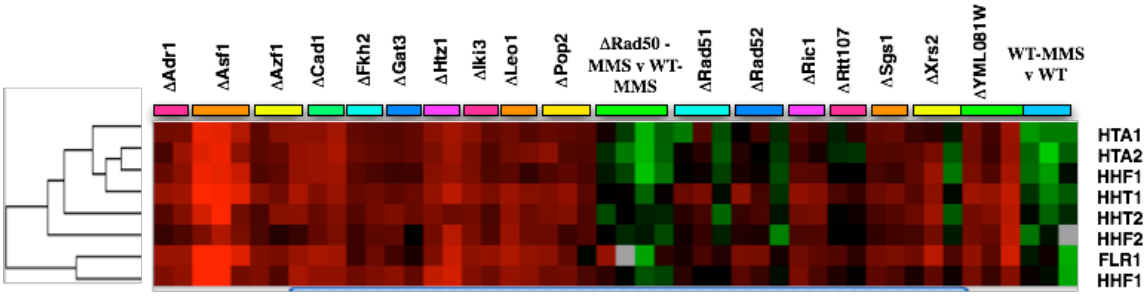
Supplemental Figure 4.1: YML081W protein similarity

Psi-BLAST of YML081W reveals that the N-terminus contains a zinc-finger domain highly conserved across many species. The C-terminal portion of the protein is similar to many fungal proteins, but few of them are characterized.



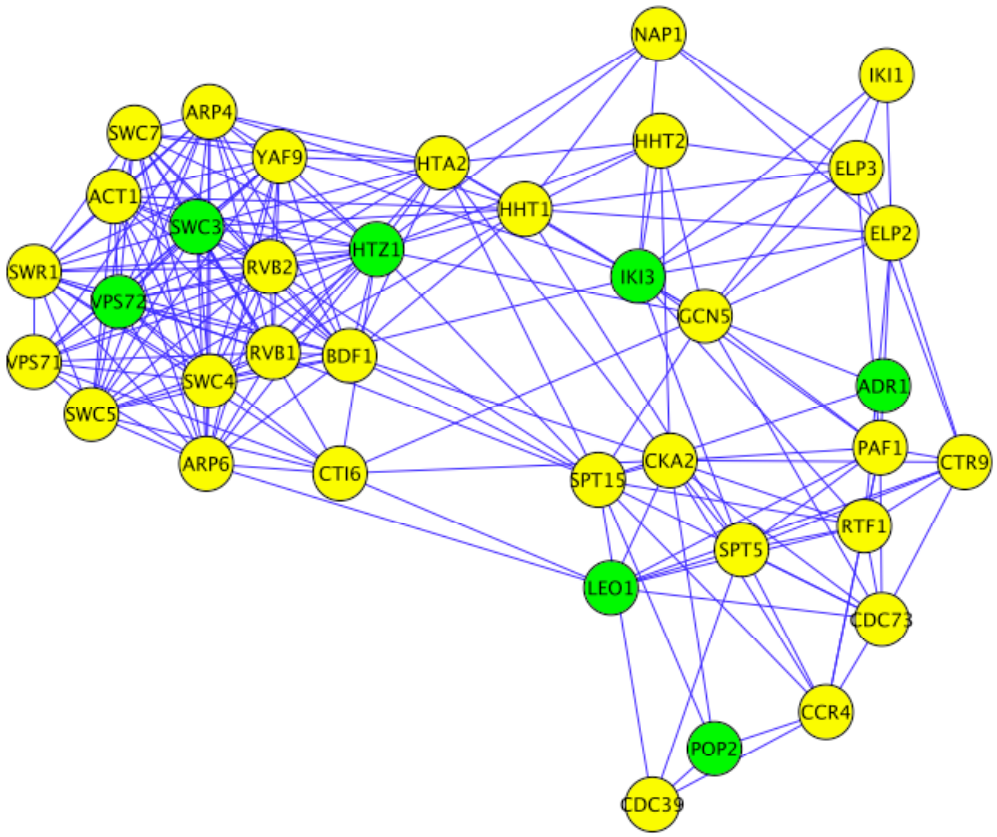
Supplemental Figure 4.2: Clustering of differentially expressed genes by Log2 ratio

A cutoff of at least two arrays with absolute value of Log2 > 2 leaves 331 microarray spots in which the strongly upregulated genes are enriched for stress response, but the most strongly downregulated genes are not.



Supplemental Figure 4.3: Differential expression of histone genes across strains

Comparison of histone transcript levels among strains shows that histone transcripts are downregulated in wild-type genes after one hour of MMS treatment (WT-MMS v WT). Levels of histone transcripts are further downregulated relative to WT-MMS in ΔRad50. Histone transcript levels in the other strains may be upregulated relative to wild type, or they may have been downregulated, but to a lesser extent than wild type.



Supplemental Figure 4.4: Leo1 and Htz1 in YeastNet 2.0

This figure shows the orientation of YML081W targets Swc3 and Vps72 relative to Htz1 in the SWR1 complex. Leo1, which is a part of the PAF complex, and Iki3, which is a member of the Elongator complex, are also shown, as are transcriptional profiling targets Pop2 and Adr1.

Chapter 5: Summary and Future Directions

While much is known about how double-stranded breaks are repaired, our current models do not fully explain the sensitivity of many mutants to cross-linking agents such MMS and mitomycin C, or double-stranded break-causing agents such as bleomycin. Because the majority of cellular DNA repair machinery is highly conserved in eukaryotes from yeast to humans, and because cancer susceptibility is so frequently a consequence of defects in DNA damage repair, it is important to understand the functions of genes involved in this process.

In order better understand the process of DNA damage repair in yeast, we screened a library of yeast deletion strains for MMS sensitivity, performed transcription profiling under DNA damage conditions, used chromatin immunoprecipitation to determine genomic loci of protein binding, and used motif analysis tools to search for transcription factor consensus motifs. We report previously undescribed MMS sensitivity for eight deletion strains, describe the unique transcriptional profiles of ΔIki3 and ΔRad50 strains under MMS treatment, and propose functional roles for Iki3 and YML081W in DNA repair. Specifically, we have found ΔHtz1 to be MMS-sensitive, and our transcriptional profiling and ChIP studies suggest a function for ΔHtz1 and SWR1 in DNA damage repair. This is supported by other studies demonstrating sensitivity of ΔHtz1, ΔVps72, ΔSwc3, ΔLeo1, and ΔYML081W to DNA-damaging compounds such as mitomycin C, hydroxyurea, and cisplatin [193]. As H2AZ is conserved across eukaryotes and essential in mice [267], any role it plays in damage repair should be characterized.

DNA damage results in phosphorylation of the histone H2A variant H2AX at C-terminal serines [267]. Yeast does not have a specific H2AX, so H2A is phosphorylated.

This phophorylation is followed by chromatin remodeling by Ino80 to facilitate damage repair [184].

H2AZ has been less studied than H2AX. It is known to be substituted for H2A by the SWR1 complex [208]; this substitution is more frequent in promoters than in coding regions and correlates with reduced transcription. It lacks the C-terminal SQ motif at which H2AX is phosphorylated [267], but H2AZ in *Tetrahymena* has essential conserved lysines in the N-terminal tail which are acetylated [268]. Interestingly, one study noted that H2Av in *Drosophila*, which is a member of the H2AZ family, is rapidly phosphorylated specifically after radiation-induced DNA damage, and imaginal disc cells from mutants with H2Av that could not be phosphorylated were more prone to apoptosis, suggesting that H2Av phosphorylation was necessary for repair of double-stranded DNA breaks [269]. However, this H2AZ homolog contains a C-terminal SQ motif [269], whereas the yeast H2AZ does not [267].

Recent studies have determined by mass spectroscopy that Htz1 in yeast is acetylated at the N terminus, and that unmodified Htz1 is associated with repressed genes, while Htz1 acetylated at K14 is associated with active genes [270]. We recommend mass spectroscopy of Htz1 after exposure to MMS to determine whether DNA damage causes modifications such as phosophorylation. Additionally, we recommend chromatin immunoprecipitation of Htz1 under normal and DNA damage conditions to assess changes in genomic distribution.

We recommend further experiments to assess the proposed role of YML081W in regulation of chromatin remodelers. First, we recommend immunoprecipitation of YML081W followed by mass spectroscopy to determine co-factors. Second, we recommend comparison of mRNA and protein levels in selected damage-related YML081W target genes in deletion and overexpression strains by Northern and Western

blot. Finally, we recommend rescue of MMS sensitivity in ΔYML081W strains by overexpression of damage-related target genes.

REFERENCES

1. Deininger PL, Batzer MA: **Mammalian retroelements.** *Genome Res* 2002, **12:**1455-1465.
2. Wray GA, Hahn MW, Abouheif E, Balhoff JP, Pizer M, Rockman MV, Romano LA: **The evolution of transcriptional regulation in eukaryotes.** *Mol Biol Evol* 2003, **20:**1377-1419.
3. Kornberg RD: **Eukaryotic transcriptional control.** *Trends Cell Biol* 1999, **9:**M46-49.
4. Hampsey M, Reinberg D: **Tails of intrigue: phosphorylation of RNA polymerase II mediates histone methylation.** *Cell* 2003, **113:**429-432.
5. Yamamoto Y, Gaynor RB: **IkappaB kinases: key regulators of the NF-kappaB pathway.** *Trends Biochem Sci* 2004, **29:**72-79.
6. King MC, Wilson AC: **Evolution at two levels in humans and chimpanzees.** *Science* 1975, **188:**107-116.
7. Gilad Y, Oshlack A, Smyth GK, Speed TP, White KP: **Expression profiling in primates reveals a rapid evolution of human transcription factors.** *Nature* 2006, **440:**242-245.
8. Fodor SP, Read JL, Pirrung MC, Stryer L, Lu AT, Solas D: **Light-directed, spatially addressable parallel chemical synthesis.** *Science* 1991, **251:**767-773.
9. Schena M, Shalon D, Davis RW, Brown PO: **Quantitative monitoring of gene expression patterns with a complementary DNA microarray.** *Science* 1995, **270:**467-470.
10. Lashkari DA, DeRisi JL, McCusker JH, Namath AF, Gentile C, Hwang SY, Brown PO, Davis RW: **Yeast microarrays for genome wide parallel genetic and gene expression analysis.** *Proc Natl Acad Sci U S A* 1997, **94:**13057-13062.
11. Wang T, Hopkins D, Schmidt C, Silva S, Houghton R, Takita H, Repasky E, Reed SG: **Identification of genes differentially over-expressed in lung squamous cell carcinoma using combination of cDNA subtraction and microarray analysis.** *Oncogene* 2000, **19:**1519-1528.
12. Xu J, Stolk JA, Zhang X, Silva SJ, Houghton RL, Matsumura M, Vedvick TS, Leslie KB, Badaro R, Reed SG: **Identification of differentially expressed genes in human prostate cancer using subtraction and microarray.** *Cancer Res* 2000, **60:**1677-1682.
13. Ishkanian AS, Malloff CA, Watson SK, DeLeeuw RJ, Chi B, Coe BP, Snijders A, Albertson DG, Pinkel D, Marra MA, et al: **A tiling resolution DNA microarray with complete coverage of the human genome.** *Nat Genet* 2004, **36:**299-303.

14. Gu J: **Exploring the global gene expression program and regulation in the response of quiescent human fibroblasts to distinct proliferative stimuli.** The University of Texas at Austin, Molecular Biology; 2005.
15. Iyer VR, Horak CE, Scafe CS, Botstein D, Snyder M, Brown PO: **Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF.** *Nature* 2001, **409:**533-538.
16. Hahn JS, Hu Z, Thiele DJ, Iyer VR: **Genome-wide analysis of the biology of stress responses through heat shock transcription factor.** *Mol Cell Biol* 2004, **24:**5249-5256.
17. Harbison CT, Gordon DB, Lee TI, Rinaldi NJ, Macisaac KD, Danford TW, Hannett NM, Tagne JB, Reynolds DB, Yoo J, et al: **Transcriptional regulatory code of a eukaryotic genome.** *Nature* 2004, **431:**99-104.
18. Morgan XC, Ni S, Miranker DP, Iyer VR: **Predicting combinatorial binding of transcription factors to regulatory elements in the human genome by association rule mining.** *BMC Bioinformatics* 2007, **8:**445.
19. Tang H, Veldman MB, Goldman D: **Characterization of a muscle-specific enhancer in human MuSK promoter reveals the essential role of myogenin in controlling activity-dependent gene regulation.** *J Biol Chem* 2006, **281:**3943-3953.
20. Shah R, Rahaman B, Hurley CK, Posch PE: **Allelic diversity in the TGFB1 regulatory region: characterization of novel functional single nucleotide polymorphisms.** *Hum Genet* 2006, **119:**61-74.
21. Kammandel B, Chowdhury K, Stoykova A, Aparicio S, Brenner S, Gruss P: **Distinct cis-essential modules direct the time-space pattern of the Pax6 gene activity.** *Dev Biol* 1999, **205:**79-97.
22. Davidson EH: *Genomic regulatory systems : development and evolution*. San Diego: Academic Press; 2001.
23. Lenhard B, Sandelin A, Mendoza L, Engstrom P, Jareborg N, Wasserman WW: **Identification of conserved regulatory elements by comparative genome analysis.** *J Biol* 2003, **2:**13.
24. Kim J, Bhinge AA, Morgan XC, Iyer VR: **Mapping DNA-protein interactions in large genomes by sequence tag analysis of genomic enrichment.** *Nat Methods* 2005, **2:**47-53.
25. Lee TI, Rinaldi NJ, Robert F, Odom DT, Bar-Joseph Z, Gerber GK, Hannett NM, Harbison CT, Thompson CM, Simon I, et al: **Transcriptional regulatory networks in Saccharomyces cerevisiae.** *Science* 2002, **298:**799-804.
26. Carroll JS, Meyer CA, Song J, Li W, Geistlinger TR, Eeckhoute J, Brodsky AS, Keeton EK, Fertuck KC, Hall GF, et al: **Genome-wide analysis of estrogen receptor binding sites.** *Nat Genet* 2006, **38:**1289-1297.
27. Zheng Y, Josefowicz SZ, Kas A, Chu TT, Gavin MA, Rudensky AY: **Genome-wide analysis of Foxp3 target genes in developing and mature regulatory T cells.** *Nature* 2007, **445:**936-940.

28. Tupler R, Perini G, Green MR: **Expressing the human genome.** *Nature* 2001, **409:**832-833.
29. Messina DN, Glasscock J, Gish W, Lovett M: **An ORFeome-based analysis of human transcription factor genes and the construction of a microarray to interrogate their expression.** *Genome Res* 2004, **14:**2041-2047.
30. Small S, Blair A, Levine M: **Regulation of even-skipped stripe 2 in the Drosophila embryo.** *Embo J* 1992, **11:**4047-4057.
31. Rivera-Pomar R, Lu X, Perrimon N, Taubert H, Jackle H: **Activation of posterior gap gene expression in the Drosophila blastoderm.** *Nature* 1995, **376:**253-256.
32. Philipsen S, Talbot D, Fraser P, Grosveld F: **The beta-globin dominant control region: hypersensitive site 2.** *Embo J* 1990, **9:**2159-2167.
33. Rothenberg EV, Ward SB: **A dynamic assembly of diverse transcription factors integrates activation and cell-type information for interleukin 2 gene regulation.** *Proc Natl Acad Sci U S A* 1996, **93:**9358-9365.
34. Crowley EM, Roeder K, Bina M: **A statistical model for locating regulatory regions in genomic DNA.** *J Mol Biol* 1997, **268:**8-14.
35. Wagner A: **A computational genomics approach to the identification of gene networks.** *Nucleic Acids Res* 1997, **25:**3594-3604.
36. Frith MC, Hansen U, Weng Z: **Detection of cis-element clusters in higher eukaryotic DNA.** *Bioinformatics* 2001, **17:**878-889.
37. Markstein M, Markstein P, Markstein V, Levine MS: **Genome-wide analysis of clustered Dorsal binding sites identifies putative target genes in the Drosophila embryo.** *Proc Natl Acad Sci U S A* 2002, **99:**763-768.
38. Berman BP, Nibu Y, Pfeiffer BD, Tomancak P, Celniker SE, Levine M, Rubin GM, Eisen MB: **Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the Drosophila genome.** *Proc Natl Acad Sci U S A* 2002, **99:**757-762.
39. Lifanov AP, Makeev VJ, Nazina AG, Papatsenko DA: **Homotypic regulatory clusters in Drosophila.** *Genome Res* 2003, **13:**579-588.
40. Wasserman WW, Fickett JW: **Identification of regulatory regions which confer muscle-specific gene expression.** *J Mol Biol* 1998, **278:**167-181.
41. Frech K, Quandt K, Werner T: **Muscle actin genes: a first step towards computational classification of tissue specific promoters.** *In Silico Biol* 1998, **1:**29-38.
42. Tronche F, Ringeisen F, Blumenfeld M, Yaniv M, Pontoglio M: **Analysis of the distribution of binding sites for a tissue-specific transcription factor in the vertebrate genome.** *J Mol Biol* 1997, **266:**231-245.
43. Kel A, Kel-Margoulis O, Babenko V, Wingender E: **Recognition of NFATp/AP-1 composite elements within genes induced upon the activation of immune cells.** *J Mol Biol* 1999, **288:**353-376.

44. Kel AE, Kel-Margoulis OV, Farnham PJ, Bartley SM, Wingender E, Zhang MQ: **Computer-assisted identification of cell cycle-related genes: new targets for E2F transcription factors.** *J Mol Biol* 2001, **309:**99-120.
45. Berman BP, Pfeiffer BD, Laverty TR, Salzberg SL, Rubin GM, Eisen MB, Celniker SE: **Computational identification of developmental enhancers: conservation and function of transcription factor binding-site clusters in Drosophila melanogaster and Drosophila pseudoobscura.** *Genome Biol* 2004, **5:**R61.
46. Halfon MS, Grad Y, Church GM, Michelson AM: **Computation-based discovery of related transcriptional regulatory modules and motifs using an experimentally validated combinatorial model.** *Genome Res* 2002, **12:**1019-1028.
47. Rebeiz M, Reeves NL, Posakony JW: **SCORE: a computational approach to the identification of cis-regulatory modules and target genes in whole-genome sequence data. Site clustering over random expectation.** *Proc Natl Acad Sci U S A* 2002, **99:**9888-9893.
48. De Bosscher K, Vanden Berghe W, Haegeman G: **The interplay between the glucocorticoid receptor and nuclear factor-kappaB or activator protein-1: molecular mechanisms for gene repression.** *Endocr Rev* 2003, **24:**488-522.
49. Bartholdy B, Matthias P: **Transcriptional control of B cell development and function.** *Gene* 2004, **327:**1-23.
50. Beer MA, Tavazoie S: **Predicting gene expression from sequence.** *Cell* 2004, **117:**185-198.
51. Das D, Banerjee N, Zhang MQ: **Interacting models of cooperative gene regulation.** *Proc Natl Acad Sci U S A* 2004, **101:**16234-16239.
52. Sharan R, Ben-Hur A, Loots GG, Ovcharenko I: **CREME: Cis-Regulatory Module Explorer for the human genome.** *Nucleic Acids Res* 2004, **32:**W253-256.
53. Brazma A, Vilo J, Ukkonen E, Valtonen K: **Data mining for regulatory elements in yeast genome.** *Proc Int Conf Intell Syst Mol Biol* 1997, **5:**65-74.
54. Horng JT, Huang HD, Jin MH, Wu LC, Huang SL: **The repetitive sequence database and mining putative regulatory elements in gene promoter regions.** *J Comput Biol* 2002, **9:**621-640.
55. Horng JT, Lin FM, Lin JH, Huang HD, Liu BJ: **Database of repetitive elements in complete genomes and data mining using transcription factor binding sites.** *IEEE Trans Inf Technol Biomed* 2003, **7:**93-100.
56. Agrawal R, and Srikant, Ramakrishnan: **Fast Algorithms for Mining Association Rules.** *Proceedings of the 20th International Conference on Very Large Data Bases, Santiago, Chile* 1994**:**487-499.
57. Pruitt KD, Tatusova T, Maglott DR: **NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins.** *Nucleic Acids Res* 2005, **33:**D501-504.

58. Yahata T, Takedatsu H, Dunwoodie SL, Braganca J, Swingler T, Withington SL, Hur J, Coser KR, Isselbacher KJ, Bhattacharya S, Shioda T: **Cloning of mouse Cited4, a member of the CITED family p300/CBP-binding transcriptional coactivators: induced expression in mammary epithelial cells.** *Genomics* 2002, **80:**601-613.
59. Kaneko M, Yang W, Matsumoto Y, Watt F, Funa K: **Activity of a novel PDGF beta-receptor enhancer during the cell cycle and upon differentiation of neuroblastoma.** *Exp Cell Res* 2006, **312:**2028-2039.
60. Chu BY, Tran K, Ku TK, Crowe DL: **Regulation of ERK1 gene expression by coactivator proteins.** *Biochem J* 2005, **392:**589-599.
61. Szlosarek PW, Balkwill FR: **Tumour necrosis factor alpha: a potential target for the therapy of solid tumours.** *Lancet Oncol* 2003, **4:**565-573.
62. Barthel R, Tsytsykova AV, Barczak AK, Tsai EY, Dascher CC, Brenner MB, Goldfeld AE: **Regulation of tumor necrosis factor alpha gene expression by mycobacteria involves the assembly of a unique enhanceosome dependent on the coactivator proteins CBP/p300.** *Mol Cell Biol* 2003, **23:**526-533.
63. Mahapatra NR, Mahata M, Ghosh S, Gayen JR, O'Connor DT, Mahata SK: **Molecular basis of neuroendocrine cell type-specific expression of the chromogranin B gene: Crucial role of the transcription factors CREB, AP-2, Egr-1 and Sp1.** *J Neurochem* 2006, **99:**119-133.
64. Wong DL, Siddall BJ, Ebert SN, Bell RA, Her S: **Phenylethanolamine N-methyltransferase gene expression: synergistic activation by Egr-1, AP-2 and the glucocorticoid receptor.** *Brain Res Mol Brain Res* 1998, **61:**154-161.
65. Tsai EY, Falvo JV, Tsytsykova AV, Barczak AK, Reimold AM, Glimcher LH, Fenton MJ, Gordon DC, Dunn IF, Goldfeld AE: **A lipopolysaccharide-specific enhancer complex involving Ets, Elk-1, Sp1, and CREB binding protein and p300 is recruited to the tumor necrosis factor alpha promoter in vivo.** *Mol Cell Biol* 2000, **20:**6084-6094.
66. Hube F, Reverdiau P, Iochmann S, Cherpi-Antar C, Gruel Y: **Characterization and functional analysis of TFPI-2 gene promoter in a human choriocarcinoma cell line.** *Thromb Res* 2003, **109:**207-215.
67. Minc E, de Coppet P, Masson P, Thiery L, Dutertre S, Amor-Gueret M, Jaulin C: **The human copper-zinc superoxide dismutase gene (SOD1) proximal promoter is regulated by Sp1, Egr-1, and WT1 via non-canonical binding sites.** *J Biol Chem* 1999, **274:**503-509.
68. Li-Weber M, Krammer PH: **Function and regulation of the CD95 (APO-1/Fas) ligand in the immune system.** *Semin Immunol* 2003, **15:**145-157.
69. Pazdrak K, Shi XZ, Sarna SK: **TNFalpha suppresses human colonic circular smooth muscle cell contractility by SP1- and NF-kappaB-mediated induction of ICAM-1.** *Gastroenterology* 2004, **127:**1096-1109.
70. Gorgoulis VG, Zacharatos P, Kotsinas A, Kletsas D, Mariatos G, Zoumpourlis V, Ryan KM, Kittas C, Papavassiliou AG: **p53 activates ICAM-1 (CD54)**

expression in an NF-kappaB-independent manner. *Embo J* 2003, **22:**1567-1578.

71. Le Mee S, Fromigue O, Marie PJ: **Sp1/Sp3 and the myeloid zinc finger gene MZF1 regulate the human N-cadherin promoter in osteoblasts.** *Exp Cell Res* 2005, **302:**129-142.
72. Sekar N, Veldhuis JD: **Involvement of Sp1 and SREBP-1a in transcriptional activation of the LDL receptor gene by insulin and LH in cultured porcine granulosa-luteal cells.** *Am J Physiol Endocrinol Metab* 2004, **287:**E128-135.
73. Schweizer M, Roder K, Zhang L, Wolf SS: **Transcription factors acting on the promoter of the rat fatty acid synthase gene.** *Biochem Soc Trans* 2002, **30:**1070-1072.
74. Kim JC, Yoon JB, Koo HS, Chung IK: **Cloning and characterization of the 5'-flanking region for the human topoisomerase III gene.** *J Biol Chem* 1998, **273:**26130-26137.
75. Herzog B, Waltner-Law M, Scott DK, Eschrich K, Granner DK: **Characterization of the human liver fructose-1,6-bisphosphatase gene promoter.** *Biochem J* 2000, **351 Pt 2:**385-392.
76. Cockerill PN, Osborne CS, Bert AG, Grotto RJ: **Regulation of GM-CSF gene transcription by core-binding factor.** *Cell Growth Differ* 1996, **7:**917-922.
77. Wolf SS, Roder K, Sickinger S, Schweizer M: **The FIRE3-mediated sterol response of the FAS promoter requires NF-Y/CBF as a coactivator.** *Biol Chem* 2001, **382:**1083-1088.
78. La Ferla K, Reimann C, Jelkmann W, Hellwig-Burgel T: **Inhibition of erythropoietin gene expression signaling involves the transcription factors GATA-2 and NF-kappaB.** *Faseb J* 2002, **16:**1811-1813.
79. Dryer RL, Covey LR: **A novel NF-kappa B-regulated site within the human I gamma 1 promoter requires p300 for optimal transcriptional activity.** *J Immunol* 2005, **175:**4499-4507.
80. Herrmann F, Trowsdale J, Huber C, Seliger B: **Cloning and functional analyses of the mouse tapasin promoter.** *Immunogenetics* 2003, **55:**379-388.
81. Maitra S, Atchison M: **BSAP can repress enhancer activity by targeting PU.1 function.** *Mol Cell Biol* 2000, **20:**1911-1922.
82. Faggioli L, Costanzo C, Donadelli M, Palmieri M: **Activation of the Interleukin-6 promoter by a dominant negative mutant of c-Jun.** *Biochim Biophys Acta* 2004, **1692:**17-24.
83. Wickremasinghe MI, Thomas LH, O'Kane CM, Uddin J, Friedland JS: **Transcriptional mechanisms regulating alveolar epithelial cell-specific CCL5 secretion in pulmonary tuberculosis.** *J Biol Chem* 2004, **279:**27199-27210.
84. Shannon MF, Coles LS, Vadas MA, Cockerill PN: **Signals for activation of the GM-CSF promoter and enhancer in T cells.** *Crit Rev Immunol* 1997, **17:**301-323.

85. The R Development Core Team: **"R: A Language and Environment for Statistical Computing."** *R Foundation for Statistical Computing, Vienna, Austria* 2005.
86. Gentleman R: *Bioinformatics and computational biology solutions using R and Bioconductor*. New York: Springer Science+Business Media; 2005.
87. Elemento O, Tavazoie S: **Fast and systematic genome-wide discovery of conserved regulatory elements using a non-alignment based approach.** *Genome Biol* 2005, **6:**R18.
88. Ramani AK, Bunescu RC, Mooney RJ, Marcotte EM: **Consolidating the set of known human protein-protein interactions in preparation for large-scale mapping of the human interactome.** *Genome Biol* 2005, **6:**R40.
89. Lehner B, Fraser AG: **A first-draft human protein-interaction map.** *Genome Biol* 2004, **5:**R63.
90. Rhodes DR, Tomlins SA, Varambally S, Mahavisno V, Barrette T, Kalyana-Sundaram S, Ghosh D, Pandey A, Chinnaiyan AM: **Probabilistic model of the human protein-protein interaction network.** *Nat Biotechnol* 2005, **23:**951-959.
91. Stelzl U, Worm U, Lalowski M, Haenig C, Brembeck FH, Goehler H, Stroedicke M, Zenkner M, Schoenherr A, Koeppen S, et al: **A human protein-protein interaction network: a resource for annotating the proteome.** *Cell* 2005, **122:**957-968.
92. Rual JF, Venkatesan K, Hao T, Hirozane-Kishikawa T, Dricot A, Li N, Berriz GF, Gibbons FD, Dreze M, Ayivi-Guedehoussou N, et al: **Towards a proteome-scale map of the human protein-protein interaction network.** *Nature* 2005, **437:**1173-1178.
93. Hart GT, Ramani AK, Marcotte EM: **How complete are current yeast and human protein-interaction networks?** *Genome Biol* 2006, **7:**120.
94. Wingender E, Dietze P, Karas H, Knuppel R: **TRANSFAC: a database on transcription factors and their DNA binding sites.** *Nucleic Acids Res* 1996, **24:**238-241.
95. Hertz GZ, Stormo GD: **Identifying DNA and protein patterns with statistically significant alignments of multiple sequences.** *Bioinformatics* 1999, **15:**563-577.
96. Karolchik D, Baertsch R, Diekhans M, Furey TS, Hinrichs A, Lu YT, Roskin KM, Schwartz M, Sugnet CW, Thomas DJ, et al: **The UCSC Genome Browser Database.** *Nucleic Acids Res* 2003, **31:**51-54.
97. Demeter J, Beauheim C, Gollub J, Hernandez-Boussard T, Jin H, Maier D, Matese JC, Nitzberg M, Wymore F, Zachariah ZK, et al: **The Stanford Microarray Database: implementation of new analysis tools and open source release of software.** *Nucleic Acids Res* 2007, **35:**D766-770.
98. Kel-Margoulis OV, Kel AE, Reuter I, Deineko IV, Wingender E: **TRANSCompel: a database on composite regulatory elements in eukaryotic genes.** *Nucleic Acids Res* 2002, **30:**332-334.

99. Bertrand-Philippe M, Ruddell RG, Arthur MJ, Thomas J, Mungalsingh N, Mann DA: **Regulation of tissue inhibitor of metalloproteinase 1 gene transcription by RUNX1 and RUNX2.** *J Biol Chem* 2004, **279:**24530-24539.
100. Maier H, Ostraat R, Gao H, Fields S, Shinton SA, Medina KL, Ikawa T, Murre C, Singh H, Hardy RR, Hagman J: **Early B cell factor cooperates with Runx1 and mediates epigenetic changes associated with mb-1 transcription.** *Nat Immunol* 2004, **5:**1069-1077.
101. Hromas R, Davis B, Rauscher FJ, 3rd, Klemsz M, Tenen D, Hoffman S, Xu D, Morris JF: **Hematopoietic transcriptional regulation by the myeloid zinc finger gene, MZF-1.** *Curr Top Microbiol Immunol* 1996, **211:**159-164.
102. Libermann TA, Pan Z, Akbarali Y, Hetherington CJ, Boltax J, Yergeau DA, Zhang DE: **AML1 (CBFalpha2) cooperates with B cell-specific activating protein (BSAP/PAX5) in activation of the B cell-specific BLK gene promoter.** *J Biol Chem* 1999, **274:**24671-24676.
103. Falvo JV, Uglialoro AM, Brinkman BM, Merika M, Parekh BS, Tsai EY, King HC, Morielli AD, Peralta EG, Maniatis T, et al: **Stimulus-specific assembly of enhancer complexes on the tumor necrosis factor alpha gene promoter.** *Mol Cell Biol* 2000, **20:**2239-2247.
104. Andriamanalijaona R, Felisaz N, Kim SJ, King-Jones K, Lehmann M, Pujol JP, Boumediene K: **Mediation of interleukin-1beta-induced transforming growth factor beta1 expression by activator protein 4 transcription factor in primary cultures of bovine articular chondrocytes: possible cooperation with activator protein 1.** *Arthritis Rheum* 2003, **48:**1569-1581.
105. Cohn MA, Hjelmso I, Wu LC, Guldberg P, Lukanidin EM, Tulchinsky EM: **Characterization of Sp1, AP-1, CBF and KRC binding sites and minisatellite DNA as functional elements of the metastasis-associated mts1/S100A4 gene intronic enhancer.** *Nucleic Acids Res* 2001, **29:**3335-3346.
106. Johnson BV, Bert AG, Ryan GR, Condina A, Cockerill PN: **Granulocyte-macrophage colony-stimulating factor enhancer activation requires cooperation between NFAT and AP-1 elements and is associated with extensive nucleosome reorganization.** *Mol Cell Biol* 2004, **24:**7914-7930.
107. Britos-Bray M, Friedman AD: **Core binding factor cannot synergistically activate the myeloperoxidase proximal enhancer in immature myeloid cells without c-Myb.** *Mol Cell Biol* 1997, **17:**5127-5135.
108. Debieve F, Thomas K: **Control of the human inhibin alpha chain promoter in cytotrophoblast cells differentiating into syncytium.** *Mol Hum Reprod* 2002, **8:**262-270.
109. Ebert SN, Ficklin MB, Her S, Siddall BJ, Bell RA, Ganguly K, Morita K, Wong DL: **Glucocorticoid-dependent action of neural crest factor AP-2: stimulation of phenylethanolamine N-methyltransferase gene expression.** *J Neurochem* 1998, **70:**2286-2295.

110. Zhou L, Nazarian AA, Smale ST: **Interleukin-10 inhibits interleukin-12 p40 gene transcription by targeting a late event in the activation pathway.** *Mol Cell Biol* 2004, **24:**2385-2396.
111. Zhou T, Chiang CM: **Sp1 and AP2 regulate but do not constitute TATA-less human TAF(II)55 core promoter activity.** *Nucleic Acids Res* 2002, **30:**4145-4157.
112. Yang H, Wang J, Ou X, Huang ZZ, Lu SC: **Cloning and analysis of the rat glutamate-cysteine ligase modifier subunit promoter.** *Biochem Biophys Res Commun* 2001, **285:**476-482.
113. Moon SK, Cha BY, Kim CH: **ERK1/2 mediates TNF-alpha-induced matrix metalloproteinase-9 expression in human vascular smooth muscle cells via the regulation of NF-kappaB and AP-1: Involvement of the ras dependent pathway.** *J Cell Physiol* 2004, **198:**417-427.
114. Shi Q, Le X, Abbruzzese JL, Wang B, Mujaida N, Matsushima K, Huang S, Xiong Q, Xie K: **Cooperation between transcription factor AP-1 and NF-kappaB in the induction of interleukin-8 in human pancreatic adenocarcinoma cells by hypoxia.** *J Interferon Cytokine Res* 1999, **19:**1363-1371.
115. Braganca J, Eloranta JJ, Bamforth SD, Ibbitt JC, Hurst HC, Bhattacharya S: **Physical and functional interactions among AP-2 transcription factors, p300/CREB-binding protein, and CITED2.** *J Biol Chem* 2003, **278:**16021-16029.
116. Becker C, Wirtz S, Ma X, Blessing M, Galle PR, Neurath MF: **Regulation of IL-12 p40 promoter activity in primary human monocytes: roles of NF-kappaB, CCAAT/enhancer-binding protein beta, and PU.1 and identification of a novel repressor element (GA-12) that responds to IL-4 and prostaglandin E(2).** *J Immunol* 2001, **167:**2608-2618.
117. Lavrovsky Y, Schwartzman ML, Levere RD, Kappas A, Abraham NG: **Identification of binding sites for transcription factors NF-kappa B and AP-2 in the promoter region of the human heme oxygenase 1 gene.** *Proc Natl Acad Sci U S A* 1994, **91:**5987-5991.
118. Barski OA, Papusha VZ, Kunkel GR, Gabbay KH: **Regulation of aldehyde reductase expression by STAF and CHOP.** *Genomics* 2004, **83:**119-129.
119. Mura C, Le Gac G, Jacolot S, Ferec C: **Transcriptional regulation of the human HFE gene indicates high liver expression and erythropoiesis coregulation.** *Faseb J* 2004, **18:**1922-1924.
120. Lahlil R, Lecuyer E, Herblot S, Hoang T: **SCL assembles a multifactorial complex that determines glycophorin A expression.** *Mol Cell Biol* 2004, **24:**1439-1452.
121. Malakooti J, Memark VC, Dudeja PK, Ramaswamy K: **Molecular cloning and functional analysis of the human Na(+)/H(+) exchanger NHE3 promoter.** *Am J Physiol Gastrointest Liver Physiol* 2002, **282:**G491-500.

122. Lin CS, Chow S, Lau A, Tu R, Lue TF: **Identification and regulation of human PDE5A gene promoter.** *Biochem Biophys Res Commun* 2001, **280:**684-692.
123. Holzmann C, Schmidt T, Thiel G, Epplen JT, Riess O: **Functional characterization of the human Huntington's disease gene promoter.** *Brain Res Mol Brain Res* 2001, **92:**85-97.
124. Gu JM, Fukudome K, Esmon CT: **Characterization and regulation of the 5'-flanking region of the murine endothelial protein C receptor gene.** *J Biol Chem* 2000, **275:**12481-12488.
125. Pocock J, Gomez-Guerrero C, Harendza S, Ayoub M, Hernandez-Vargas P, Zahner G, Stahl RA, Thaiss F: **Differential activation of NF-kappa B, AP-1, and C/EBP in endotoxin-tolerant rats: mechanisms for in vivo regulation of glomerular RANTES/CCL5 expression.** *J Immunol* 2003, **170:**6280-6291.
126. Rojo AI, Salinas M, Martin D, Perona R, Cuadrado A: **Regulation of Cu/Zn-superoxide dismutase expression via the phosphatidylinositol 3 kinase/Akt pathway and nuclear factor-kappaB.** *J Neurosci* 2004, **24:**7324-7334.
127. Seo SJ, Kim HT, Cho G, Rho HM, Jung G: **Sp1 and C/EBP-related factor regulate the transcription of human Cu/Zn SOD gene.** *Gene* 1996, **178:**177-185.
128. Kim HT, Kim YH, Nam JW, Lee HJ, Rho HM, Jung G: **Study of 5'-flanking region of human Cu/Zn superoxide dismutase.** *Biochem Biophys Res Commun* 1994, **201:**1526-1533.
129. Xu Z, Dziarski R, Wang Q, Swartz K, Sakamoto KM, Gupta D: **Bacterial peptidoglycan-induced tnf-alpha transcription is mediated through the transcription factors Egr-1, Elk-1, and NF-kappaB.** *J Immunol* 2001, **167:**6975-6982.
130. Yu X, Zhu X, Pi W, Ling J, Ko L, Takeda Y, Tuan D: **The long terminal repeat (LTR) of ERV-9 human endogenous retrovirus binds to NF-Y in the assembly of an active LTR enhancer complex NF-Y/MZF1/GATA-2.** *J Biol Chem* 2005, **280:**35184-35194.
131. Han L, Lu J, Pan L, Wang X, Shao Y, Han S, Huang B: **Histone acetyltransferase p300 regulates the transcription of human erythroid-specific 5-aminolevulinate synthase gene.** *Biochem Biophys Res Commun* 2006, **348:**799-806.
132. Neish AS, Williams AJ, Palmer HJ, Whitley MZ, Collins T: **Functional analysis of the human vascular cell adhesion molecule 1 promoter.** *J Exp Med* 1992, **176:**1583-1593.
133. Da Silva CA, Heilbock C, Kassel O, Frossard N: **Transcription of stem cell factor (SCF) is potentiated by glucocorticoids and interleukin-1beta through concerted regulation of a GRE-like and an NF-kappaB response element.** *Faseb J* 2003, **17:**2334-2336.
134. Hermoso MA, Matsuguchi T, Smoak K, Cidlowski JA: **Glucocorticoids and tumor necrosis factor alpha cooperatively regulate toll-like receptor 2 gene expression.** *Mol Cell Biol* 2004, **24:**4743-4756.

135. Khan S, Barhoumi R, Burghardt R, Liu S, Kim K, Safe S: **Molecular mechanism of inhibitory aryl hydrocarbon receptor-estrogen receptor/Sp1 cross talk in breast cancer cells.** *Mol Endocrinol* 2006, **20:**2199-2214.
136. Manoli I, Le H, Alesci S, McFann KK, Su YA, Kino T, Chrousos GP, Blackman MR: **Monoamine oxidase-A is a major target gene for glucocorticoids in human skeletal muscle cells.** *Faseb J* 2005, **19:**1359-1361.
137. Gobin SJ, Biesta P, Van den Elsen PJ: **Regulation of human beta 2-microglobulin transactivation in hematopoietic cells.** *Blood* 2003, **101:**3058-3064.
138. Wu CX, Zhao WP, Kishi H, Dokan J, Jin ZX, Wei XC, Yokoyama KK, Muraguchi A: **Activation of mouse RAG-2 promoter by Myc-associated zinc finger protein.** *Biochem Biophys Res Commun* 2004, **317:**1096-1102.
139. Biesiada E, Hamamori Y, Kedes L, Sartorelli V: **Myogenic basic helix-loop-helix proteins and Sp1 interact as components of a multiprotein transcriptional complex required for activity of the human cardiac alpha-actin promoter.** *Mol Cell Biol* 1999, **19:**2577-2584.
140. Kang NY, Park YD, Choi HJ, Kim KS, Lee YC, Kim CH: **Regulatory elements involved in transcription of the human NeuAcalpha2,3Galbeta1,3GalNAcalpha2,6-sialyltransferase (hST6GalNAc IV) gene.** *Mol Cells* 2004, **18:**157-162.
141. Furlong EE, Rein T, Martin F: **YY1 and NF1 both activate the human p53 promoter by alternatively binding to a composite element, and YY1 and E1A cooperate to amplify p53 promoter activity.** *Mol Cell Biol* 1996, **16:**5933-5945.
142. Inoue A, Omoto Y, Yamaguchi Y, Kiyama R, Hayashi SI: **Transcription factor EGR3 is involved in the estrogen-signaling pathway in breast cancer cells.** *J Mol Endocrinol* 2004, **32:**649-661.
143. Xiao S, Marshak-Rothstein A, Ju ST: **Sp1 is the major fasl gene activator in abnormal CD4(-)CD8(-)B220(+) T cells of lpr and gld mice.** *Eur J Immunol* 2001, **31:**3339-3348.
144. Golubovskaya V, Kaur A, Cance W: **Cloning and characterization of the promoter region of human focal adhesion kinase gene: nuclear factor kappa B and p53 binding sites.** *Biochim Biophys Acta* 2004, **1678:**111-125.
145. Schafer H, Diebel J, Arlt A, Trauzold A, Schmidt WE: **The promoter of human p22/PACAP response gene 1 (PRG1) contains functional binding sites for the p53 tumor suppressor and for NFkappaB.** *FEBS Lett* 1998, **436:**139-143.
146. Hoffmeister A, Ropolo A, Vasseur S, Mallo GV, Bodeker H, Ritz-Laser B, Dressler GR, Vaccaro MI, Dagorn JC, Moreno S, Iovanna JL: **The HMG-I/Y-related protein p8 binds to p300 and Pax2 trans-activation domain-interacting protein to regulate the trans-activation activity of the Pax2A and Pax2B transcription factors on the glucagon gene promoter.** *J Biol Chem* 2002, **277:**22314-22319.

147. Gordon SJ, Saleque S, Birshtein BK: **Yin Yang 1 is a lipopolysaccharide-inducible activator of the murine 3' Igh enhancer, hs3.** *J Immunol* 2003, **170:**5549-5557.
148. Ikeda Y, Yamamoto J, Okamura M, Fujino T, Takahashi S, Takeuchi K, Osborne TF, Yamamoto TT, Ito S, Sakai J: **Transcriptional regulation of the murine acetyl-CoA synthetase 1 gene through multiple clustered binding sites for sterol regulatory element-binding proteins and a single neighboring site for Sp1.** *J Biol Chem* 2001, **276:**34259-34269.
149. Armelin-Correa LM, Lin CJ, Barbosa A, Bagatini K, Winnischofer SM, Sogayar MC, Passos-Bueno MR: **Characterization of human Collagen XVIII promoter 2: interaction of Sp1, Sp3 and YY1 with the regulatory region and a SNP that increases transcription in hepatocytes.** *Matrix Biol* 2005, **24:**550-559.
150. Kawada H, Nishiyama C, Takagi A, Tokura T, Nakano N, Maeda K, Mayuzumi N, Ikeda S, Okumura K, Ogawa H: **Transcriptional regulation of ATP2C1 gene by Sp1 and YY1 and reduced function of its promoter in Hailey-Hailey disease keratinocytes.** *J Invest Dermatol* 2005, **124:**1206-1214.
151. Perrotti D, Melotti P, Skorski T, Casella I, Peschle C, Calabretta B: **Overexpression of the zinc finger protein MZF1 inhibits hematopoietic development from embryonic stem cells: correlation with negative regulation of CD34 and c-myb promoter activity.** *Mol Cell Biol* 1995, **15:**6075-6087.
152. Iyer VR: *Microarray-based detection of DNA protein interactions: Chromatin Immunoprecipitation on microarrays*. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press; 2003.
153. Shalon D: **DNA Micro Arrays: A New Tool for Genetic Analysis.** 1995.
154. DeRisi JL, Iyer VR, Brown PO: **Exploring the metabolic and genetic control of gene expression on a genomic scale.** *Science* 1997, **278:**680-686.
155. Killion PJ, Sherlock G, Iyer VR: **The Longhorn Array Database (LAD): an open-source, MIAME compliant implementation of the Stanford Microarray Database (SMD).** *BMC Bioinformatics* 2003, **4:**32.
156. Killion PJ, Iyer VR: **Microarray data visualization and analysis with the Longhorn Array Database (LAD).** *Curr Protoc Bioinformatics* 2004, **Chapter 7:**Unit 7 10.
157. Hu Z: **Functional Transcription Regulatory Network Reconstruction and Characterization.** The University of Texas at Austin, Microbiology; 2005.
158. Hu Z, Killion PJ, Iyer VR: **Genetic reconstruction of a functional transcriptional regulatory network.** *Nat Genet* 2007, **39:**683-687.
159. Hughes TR, Marton MJ, Jones AR, Roberts CJ, Stoughton R, Armour CD, Bennett HA, Coffey E, Dai H, He YD, et al: **Functional discovery via a compendium of expression profiles.** *Cell* 2000, **102:**109-126.
160. Winzeler EA, Shoemaker DD, Astromoff A, Liang H, Anderson K, Andre B, Bangham R, Benito R, Boeke JD, Bussey H, et al: **Functional characterization of the S. cerevisiae genome by gene deletion and parallel analysis.** *Science* 1999, **285:**901-906.

161. Baudin A, Ozier-Kalogeropoulos O, Denouel A, Lacroute F, Cullin C: **A simple and efficient method for direct gene deletion in Saccharomyces cerevisiae.** *Nucleic Acids Res* 1993, **21:**3329-3330.
162. Wach A, Brachat A, Pohlmann R, Philippsen P: **New heterologous modules for classical or PCR-based gene disruptions in Saccharomyces cerevisiae.** *Yeast* 1994, **10:**1793-1808.
163. Hoffman CS: **Preparation of yeast DNA.** *Curr Protoc Mol Biol* 2001, **Chapter 13:**Unit13 11.
164. Kim H, Chen J: **New Players in the BRCA1-mediated DNA Damage Responsive Pathway.** *Mol Cells* 2008, **25:**457-461.
165. Shiloh Y, Kastan MB: **ATM: genome stability, neuronal development, and cancer cross paths.** *Adv Cancer Res* 2001, **83:**209-254.
166. Nouspikel T: **Nucleotide excision repair and neurological diseases.** *DNA Repair (Amst)* 2008, **7:**1155-1167.
167. Shrivastav M, De Haro LP, Nickoloff JA: **Regulation of DNA double-strand break repair pathway choice.** *Cell Res* 2008, **18:**134-147.
168. Lisby M, Barlow JH, Burgess RC, Rothstein R: **Choreography of the DNA damage response: spatiotemporal relationships among checkpoint and repair proteins.** *Cell* 2004, **118:**699-713.
169. Lukas J, Bartek J: **Watching the DNA repair ensemble dance.** *Cell* 2004, **118:**666-668.
170. Paull TT, Gellert M: **The 3' to 5' exonuclease activity of Mre 11 facilitates repair of DNA double-strand breaks.** *Mol Cell* 1998, **1:**969-979.
171. Furuse M, Nagase Y, Tsubouchi H, Murakami-Murofushi K, Shibata T, Ohta K: **Distinct roles of two separable in vitro activities of yeast Mre11 in mitotic and meiotic recombination.** *Embo J* 1998, **17:**6412-6425.
172. Anderson DE, Trujillo KM, Sung P, Erickson HP: **Structure of the Rad50 x Mre11 DNA repair complex from Saccharomyces cerevisiae by electron microscopy.** *J Biol Chem* 2001, **276:**37027-37033.
173. Raymond WE, Kleckner N: **RAD50 protein of S.cerevisiae exhibits ATP-dependent DNA binding.** *Nucleic Acids Res* 1993, **21:**3851-3856.
174. Chen L, Trujillo KM, Van Komen S, Roh DH, Krejci L, Lewis LK, Resnick MA, Sung P, Tomkinson AE: **Effect of amino acid substitutions in the rad50 ATP binding domain on DNA double strand break repair in yeast.** *J Biol Chem* 2005, **280:**2620-2627.
175. Trujillo KM, Roh DH, Chen L, Van Komen S, Tomkinson A, Sung P: **Yeast xrs2 binds DNA and helps target rad50 and mre11 to DNA ends.** *J Biol Chem* 2003, **278:**48957-48964.
176. Usui T, Ohta T, Oshiumi H, Tomizawa J, Ogawa H, Ogawa T: **Complex formation and functional versatility of Mre11 of budding yeast in recombination.** *Cell* 1998, **95:**705-716.
177. Ira G, Pellicioli A, Balijja A, Wang X, Fiorani S, Carotenuto W, Liberi G, Bressan D, Wan L, Hollingsworth NM, et al: **DNA end resection, homologous**

recombination and DNA damage checkpoint activation require CDK1. *Nature* 2004, **431:**1011-1017.
178. Lisby M, Rothstein R: **Localization of checkpoint and repair proteins in eukaryotes.** *Biochimie* 2005, **87:**579-589.
179. Majka J, Niedziela-Majka A, Burgers PM: **The checkpoint clamp activates Mec1 kinase during initiation of the DNA damage checkpoint.** *Mol Cell* 2006, **24:**891-901.
180. Vialard JE, Gilbert CS, Green CM, Lowndes NF: **The budding yeast Rad9 checkpoint protein is subjected to Mec1/Tel1-dependent hyperphosphorylation and interacts with Rad53 after DNA damage.** *Embo J* 1998, **17:**5679-5688.
181. Emili A: **MEC1-dependent phosphorylation of Rad9p in response to DNA damage.** *Mol Cell* 1998, **2:**183-189.
182. Sweeney FD, Yang F, Chi A, Shabanowitz J, Hunt DF, Durocher D: **Saccharomyces cerevisiae Rad9 acts as a Mec1 adaptor to allow Rad53 activation.** *Curr Biol* 2005, **15:**1364-1375.
183. Blankley RT, Lydall D: **A domain of Rad9 specifically required for activation of Chk1 in budding yeast.** *J Cell Sci* 2004, **117:**601-608.
184. Morrison AJ, Highland J, Krogan NJ, Arbel-Eden A, Greenblatt JF, Haber JE, Shen X: **INO80 and gamma-H2AX interaction links ATP-dependent chromatin remodeling to DNA damage repair.** *Cell* 2004, **119:**767-775.
185. Morrison AJ, Kim JA, Person MD, Highland J, Xiao J, Wehr TS, Hensley S, Bao Y, Shen J, Collins SR, et al: **Mec1/Tel1 phosphorylation of the INO80 chromatin remodeling complex influences DNA damage checkpoint responses.** *Cell* 2007, **130:**499-511.
186. Denis CL, Young ET: **Isolation and characterization of the positive regulatory gene ADR1 from Saccharomyces cerevisiae.** *Mol Cell Biol* 1983, **3:**360-370.
187. Young ET, Dombek KM, Tachibana C, Ideker T: **Multiple pathways are co-regulated by the protein kinase Snf1 and the transcription factors Adr1 and Cat8.** *J Biol Chem* 2003, **278:**26146-26158.
188. Cook WJ, Chase D, Audino DC, Denis CL: **Dissection of the ADR1 protein reveals multiple, functionally redundant activation domains interspersed with inhibitory regions: evidence for a repressor binding to the ADR1c region.** *Mol Cell Biol* 1994, **14:**629-640.
189. Cheng C, Kacherovsky N, Dombek KM, Camier S, Thukral SK, Rhim E, Young ET: **Identification of potential target genes for Adr1p through characterization of essential nucleotides in UAS1.** *Mol Cell Biol* 1994, **14:**3842-3852.
190. Tachibana C, Yoo JY, Tagne JB, Kacherovsky N, Lee TI, Young ET: **Combined global localization analysis and transcriptome data identify genes that are directly coregulated by Adr1 and Cat8.** *Mol Cell Biol* 2005, **25:**2138-2146.
191. Haurie V, Perrot M, Mini T, Jeno P, Sagliocco F, Boucherie H: **The transcriptional activator Cat8p provides a major contribution to the**

reprogramming of carbon metabolism during the diauxic shift in Saccharomyces cerevisiae. *J Biol Chem* 2001, **276:**76-85.
192. Li B, Pattenden SG, Lee D, Gutierrez J, Chen J, Seidel C, Gerton J, Workman JL: **Preferential occupancy of histone variant H2AZ at inactive promoters influences local histone modifications and chromatin remodeling.** *Proc Natl Acad Sci U S A* 2005, **102:**18385-18390.
193. Hillenmeyer ME, Fung E, Wildenhain J, Pierce SE, Hoon S, Lee W, Proctor M, St Onge RP, Tyers M, Koller D, et al: **The chemical genomic portrait of yeast: uncovering a phenotype for all genes.** *Science* 2008, **320:**362-365.
194. Tyler JK, Adams CR, Chen SR, Kobayashi R, Kamakaka RT, Kadonaga JT: **The RCAF complex mediates chromatin assembly during DNA replication and repair.** *Nature* 1999, **402:**555-560.
195. Emili A, Schieltz DM, Yates JR, 3rd, Hartwell LH: **Dynamic interaction of DNA damage checkpoint protein Rad53 with chromatin assembly factor Asf1.** *Mol Cell* 2001, **7:**13-20.
196. Zappulla DC, Maharaj AS, Connelly JJ, Jockusch RA, Sternglanz R: **Rtt107/Esc4 binds silent chromatin and DNA repair proteins using different BRCT motifs.** *BMC Mol Biol* 2006, **7:**40.
197. Mousson F, Lautrette A, Thuret JY, Agez M, Courbeyrette R, Amigues B, Becker E, Neumann JM, Guerois R, Mann C, Ochsenbein F: **Structural basis for the interaction of Asf1 with histone H3 and its functional implications.** *Proc Natl Acad Sci U S A* 2005, **102:**5975-5980.
198. Ramey CJ, Howar S, Adkins M, Linger J, Spicer J, Tyler JK: **Activation of the DNA damage checkpoint in yeast lacking the histone chaperone anti-silencing function 1.** *Mol Cell Biol* 2004, **24:**10313-10327.
199. Xia L, Jaafar L, Cashikar A, Flores-Rozas H: **Identification of genes required for protection from doxorubicin by a genome-wide screen in Saccharomyces cerevisiae.** *Cancer Res* 2007, **67:**11411-11418.
200. Newcomb LL, Hall DD, Heideman W: **AZF1 is a glucose-dependent positive regulator of CLN3 transcription in Saccharomyces cerevisiae.** *Mol Cell Biol* 2002, **22:**1607-1614.
201. Kucejova B, Foury F: **Search for protein partners of mitochondrial single-stranded DNA-binding protein Rim1p using a yeast two-hybrid system.** *Folia Microbiol (Praha)* 2003, **48:**183-188.
202. Cohen BA, Pilpel Y, Mitra RD, Church GM: **Discrimination between paralogs using microarray analysis: application to the Yap1p and Yap2p transcriptional networks.** *Mol Biol Cell* 2002, **13:**1608-1614.
203. Fernandes L, Rodrigues-Pousada C, Struhl K: **Yap, a novel family of eight bZIP proteins in Saccharomyces cerevisiae with distinct biological functions.** *Mol Cell Biol* 1997, **17:**6982-6993.
204. Kumar R, Reynolds DM, Shevchenko A, Shevchenko A, Goldstone SD, Dalton S: **Forkhead transcription factors, Fkh1p and Fkh2p, collaborate with Mcm1p to control transcription required for M-phase.** *Curr Biol* 2000, **10:**896-906.

205. Kaufmann E, Muller D, Knochel W: **DNA recognition site analysis of Xenopus winged helix proteins.** *J Mol Biol* 1995, **248:**239-254.
206. Cox KH, Pinchak AB, Cooper TG: **Genome-wide transcriptional analysis in S. cerevisiae by mini-array membrane hybridization.** *Yeast* 1999, **15:**703-713.
207. McClellan AJ, Xia Y, Deutschbauer AM, Davis RW, Gerstein M, Frydman J: **Diverse cellular functions of the Hsp90 molecular chaperone uncovered using systems approaches.** *Cell* 2007, **131:**121-135.
208. Mizuguchi G, Shen X, Landry J, Wu WH, Sen S, Wu C: **ATP-driven exchange of histone H2AZ variant catalyzed by SWR1 chromatin remodeling complex.** *Science* 2004, **303:**343-348.
209. Dudley AM, Janse DM, Tanay A, Shamir R, Church GM: **A global view of pleiotropy and phenotypically derived gene function in yeast.** *Mol Syst Biol* 2005, **1:**2005 0001.
210. Wittschieben BO, Otero G, de Bizemont T, Fellows J, Erdjument-Bromage H, Ohba R, Li Y, Allis CD, Tempst P, Svejstrup JQ: **A novel histone acetyltransferase is an integral subunit of elongating RNA polymerase II holoenzyme.** *Mol Cell* 1999, **4:**123-128.
211. Winkler GS, Kristjuhan A, Erdjument-Bromage H, Tempst P, Svejstrup JQ: **Elongator is a histone H3 and H4 acetyltransferase important for normal histone acetylation levels in vivo.** *Proc Natl Acad Sci U S A* 2002, **99:**3517-3522.
212. Krogan NJ, Dover J, Wood A, Schneider J, Heidt J, Boateng MA, Dean K, Ryan OW, Golshani A, Johnston M, et al: **The Paf1 complex is required for histone H3 methylation by COMPASS and Dot1p: linking transcriptional elongation to histone methylation.** *Mol Cell* 2003, **11:**721-729.
213. Sakai A, Chibazakura T, Shimizu Y, Hishinuma F: **Molecular analysis of POP2 gene, a gene required for glucose-derepression of gene expression in Saccharomyces cerevisiae.** *Nucleic Acids Res* 1992, **20:**6227-6233.
214. Daugeron MC, Mauxion F, Seraphin B: **The yeast POP2 gene encodes a nuclease involved in mRNA deadenylation.** *Nucleic Acids Res* 2001, **29:**2448-2455.
215. Woolstencroft RN, Beilharz TH, Cook MA, Preiss T, Durocher D, Tyers M: **Ccr4 contributes to tolerance of replication stress through control of CRT1 mRNA poly(A) tail length.** *J Cell Sci* 2006, **119:**5178-5192.
216. Legrand M, Chan CL, Jauert PA, Kirkpatrick DT: **Role of DNA mismatch repair and double-strand break repair in genome stability and antifungal drug resistance in Candida albicans.** *Eukaryot Cell* 2007, **6:**2194-2205.
217. Sung P: **Catalysis of ATP-dependent homologous DNA pairing and strand exchange by yeast RAD51 protein.** *Science* 1994, **265:**1241-1243.
218. St Onge RP, Mani R, Oh J, Proctor M, Fung E, Davis RW, Nislow C, Roth FP, Giaever G: **Systematic pathway analysis using high-resolution fitness profiling of combinatorial gene deletions.** *Nat Genet* 2007, **39:**199-206.

219. Sung P: **Function of yeast Rad52 protein as a mediator between replication protein A and the Rad51 recombinase.** *J Biol Chem* 1997, **272:**28194-28197.
220. Teo SH, Jackson SP: **Identification of Saccharomyces cerevisiae DNA ligase IV: involvement in DNA double-strand break repair.** *Embo J* 1997, **16:**4788-4795.
221. Chai B, Huang J, Cairns BR, Laurent BC: **Distinct roles for the RSC and Swi/Snf ATP-dependent chromatin remodelers in DNA double-strand break repair.** *Genes Dev* 2005, **19:**1656-1661.
222. Mizuta K, Park JS, Sugiyama M, Nishiyama M, Warner JR: **RIC1, a novel gene required for ribosome synthesis in Saccharomyces cerevisiae.** *Gene* 1997, **187:**171-178.
223. Bensen ES, Yeung BG, Payne GS: **Ric1p and the Ypt6p GTPase function in a common pathway required for localization of trans-Golgi network membrane proteins.** *Mol Biol Cell* 2001, **12:**13-26.
224. Rouse J: **Esc4p, a new target of Mec1p (ATR), promotes resumption of DNA synthesis after DNA damage.** *Embo J* 2004, **23:**1188-1197.
225. Roberts TM, Kobor MS, Bastin-Shanower SA, Ii M, Horte SA, Gin JW, Emili A, Rine J, Brill SJ, Brown GW: **Slx4 regulates DNA damage checkpoint-dependent phosphorylation of the BRCT domain protein Rtt107/Esc4.** *Mol Biol Cell* 2006, **17:**539-548.
226. Bennett RJ, Sharp JA, Wang JC: **Purification and characterization of the Sgs1 DNA helicase activity of Saccharomyces cerevisiae.** *J Biol Chem* 1998, **273:**9644-9650.
227. Fabre F, Chan A, Heyer WD, Gangloff S: **Alternate pathways involving Sgs1/Top3, Mus81/ Mms4, and Srs2 prevent formation of toxic recombination intermediates from single-stranded gaps created by DNA replication.** *Proc Natl Acad Sci U S A* 2002, **99:**16887-16892.
228. Watt PM, Louis EJ, Borts RH, Hickson ID: **Sgs1: a eukaryotic homolog of E. coli RecQ that interacts with topoisomerase II in vivo and is required for faithful chromosome segregation.** *Cell* 1995, **81:**253-260.
229. Watt PM, Hickson ID, Borts RH, Louis EJ: **SGS1, a homologue of the Bloom's and Werner's syndrome genes, is required for maintenance of genome stability in Saccharomyces cerevisiae.** *Genetics* 1996, **144:**935-945.
230. Giaever G, Chu AM, Ni L, Connelly C, Riles L, Veronneau S, Dow S, Lucau-Danila A, Anderson K, Andre B, et al: **Functional profiling of the Saccharomyces cerevisiae genome.** *Nature* 2002, **418:**387-391.
231. Huh WK, Falvo JV, Gerke LC, Carroll AS, Howson RW, Weissman JS, O'Shea EK: **Global analysis of protein localization in budding yeast.** *Nature* 2003, **425:**686-691.
232. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acids Res* 1997, **25:**3389-3402.

233. Lin-Cereghino GP, Godfrey L, de la Cruz BJ, Johnson S, Khuongsathiene S, Tolstorukov I, Yan M, Lin-Cereghino J, Veenhuis M, Subramani S, Cregg JM: **Mxr1p, a key regulator of the methanol utilization pathway and peroxisomal genes in Pichia pastoris.** *Mol Cell Biol* 2006, **26:**883-897.
234. Lu L, Roberts GG, Oszust C, Hudson AP: **The YJR127C/ZMS1 gene product is involved in glycerol-based respiratory growth of the yeast Saccharomyces cerevisiae.** *Curr Genet* 2005, **48:**235-246.
235. Hong EL, Balakrishnan R, Dong Q, Christie KR, Park J, Binkley G, Costanzo MC, Dwight SS, Engel SR, Fisk DG, et al: **Gene Ontology annotations at SGD: new data sources and annotation methods.** *Nucl Acids Res* 2008, **36:**D577-581.
236. Eisen MB, Spellman PT, Brown PO, Botstein D: **Cluster analysis and display of genome-wide expression patterns.** *Proc Natl Acad Sci U S A* 1998, **95:**14863-14868.
237. **Saccharomyces Genome Database**
238. Bailey TL, Elkan C: **The value of prior knowledge in discovering motifs with MEME.** *Proc Int Conf Intell Syst Mol Biol* 1995, **3:**21-29.
239. Roth FP, Hughes JD, Estep PW, Church GM: **Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation.** *Nat Biotechnol* 1998, **16:**939-945.
240. Liu XS, Brutlag DL, Liu JS: **An algorithm for finding protein-DNA binding sites with applications to chromatin-immunoprecipitation microarray experiments.** *Nat Biotechnol* 2002, **20:**835-839.
241. Workman CT, Yin Y, Corcoran DL, Ideker T, Stormo GD, Benos PV: **enoLOGOS: a versatile web tool for energy normalized sequence logos.** *Nucleic Acids Res* 2005, **33:**W389-392.
242. Workman CT, Mak HC, McCuine S, Tagne JB, Agarwal M, Ozier O, Begley TJ, Samson LD, Ideker T: **A systems approach to mapping DNA damage response pathways.** *Science* 2006, **312:**1054-1059.
243. Ciriacy M: **Genetics of alcohol dehydrogenase in Saccharomyces cerevisiae. II. Two loci controlling synthesis of the glucose-repressible ADH II.** *Mol Gen Genet* 1975, **138:**157-164.
244. Gasch AP, Huang M, Metzner S, Botstein D, Elledge SJ, Brown PO: **Genomic expression responses to DNA-damaging agents and the regulatory role of the yeast ATR homolog Mec1p.** *Mol Biol Cell* 2001, **12:**2987-3003.
245. Martinez-Pastor MT, Marchler G, Schuller C, Marchler-Bauer A, Ruis H, Estruch F: **The Saccharomyces cerevisiae zinc finger proteins Msn2p and Msn4p are required for transcriptional induction through the stress response element (STRE).** *Embo J* 1996, **15:**2227-2235.
246. Kobayashi N, McEntee K: **Identification of cis and trans components of a novel heat shock stress regulatory pathway in Saccharomyces cerevisiae.** *Mol Cell Biol* 1993, **13:**248-256.
247. Lee I, Marcotte EM: **In preparation.** 2008.

248. Lee I, Date SV, Adai AT, Marcotte EM: **A probabilistic functional network of yeast genes.** *Science* 2004, **306:**1555-1558.
249. Lee I, Li Z, Marcotte EM: **An improved, bias-reduced probabilistic functional gene network of baker's yeast, Saccharomyces cerevisiae.** *PLoS ONE* 2007, **2:**e988.
250. Wu WH, Alami S, Luk E, Wu CH, Sen S, Mizuguchi G, Wei D, Wu C: **Swc2 is a widely conserved H2AZ-binding module essential for ATP-dependent histone exchange.** *Nat Struct Mol Biol* 2005, **12:**1064-1071.
251. Krogan NJ, Keogh MC, Datta N, Sawa C, Ryan OW, Ding H, Haw RA, Pootoolal J, Tong A, Canadien V, et al: **A Snf2 family ATPase complex required for recruitment of the histone H2A variant Htz1.** *Mol Cell* 2003, **12:**1565-1576.
252. Meeusen S, Tieu Q, Wong E, Weiss E, Schieltz D, Yates JR, Nunnari J: **Mgm101p is a novel component of the mitochondrial nucleoid that binds DNA and is required for the repair of oxidatively damaged mitochondrial DNA.** *J Cell Biol* 1999, **145:**291-304.
253. Machado CR, Praekelt UM, de Oliveira RC, Barbosa AC, Byrne KL, Meacock PA, Menck CF: **Dual role for the yeast THI4 gene in thiamine biosynthesis and DNA damage tolerance.** *J Mol Biol* 1997, **273:**114-121.
254. Torres-Ramos CA, Prakash S, Prakash L: **Requirement of RAD5 and MMS2 for postreplication repair of UV-damaged DNA in Saccharomyces cerevisiae.** *Mol Cell Biol* 2002, **22:**2419-2426.
255. Chanet R, Heude M: **Characterization of mutations that are synthetic lethal with pol3-13, a mutated allele of DNA polymerase delta in Saccharomyces cerevisiae.** *Curr Genet* 2003, **43:**337-350.
256. Lee KK, Prochasson P, Florens L, Swanson SK, Washburn MP, Workman JL: **Proteomic analysis of chromatin-modifying complexes in Saccharomyces cerevisiae identifies novel subunits.** *Biochem Soc Trans* 2004, **32:**899-903.
257. Baetz KK, Krogan NJ, Emili A, Greenblatt J, Hieter P: **The ctf13-30/CTF13 genomic haploinsufficiency modifier screen identifies the yeast chromatin remodeling complex RSC, which is required for the establishment of sister chromatid cohesion.** *Mol Cell Biol* 2004, **24:**1232-1244.
258. Nasmyth K, Dirick L: **The role of SWI4 and SWI6 in the activity of G1 cyclins in yeast.** *Cell* 1991, **66:**995-1013.
259. Spellman PT, Sherlock G, Zhang MQ, Iyer VR, Anders K, Eisen MB, Brown PO, Botstein D, Futcher B: **Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization.** *Mol Biol Cell* 1998, **9:**3273-3297.
260. Fortin GS, Symington LS: **Mutations in yeast Rad51 that partially bypass the requirement for Rad55 and Rad57 in DNA repair by increasing the stability of Rad51-DNA complexes.** *Embo J* 2002, **21:**3160-3170.
261. Aboussekhra A, Chanet R, Zgaga Z, Cassier-Chauvat C, Heude M, Fabre F: **RADH, a gene of Saccharomyces cerevisiae encoding a putative DNA**

helicase involved in DNA repair. Characteristics of radH mutants and sequence of the gene. *Nucleic Acids Res* 1989, **17:**7211-7219.
262. Rong L, Klein HL: **Purification and characterization of the SRS2 DNA helicase of the yeast Saccharomyces cerevisiae.** *J Biol Chem* 1993, **268:**1252-1259.
263. Chiolo I, Carotenuto W, Maffioletti G, Petrini JHJ, Foiani M, Liberi G: **Srs2 and Sgs1 DNA Helicases Associate with Mre11 in Different Subcomplexes following Checkpoint Activation and CDK1-Mediated Srs2 Phosphorylation.** *Mol Cell Biol* 2005, **25:**5738-5751.
264. Bradshaw V, McEntee K: **DNA damage activates the transcription and transposition of yeast TY retrotransposons.** *Molecular and General Genetics* 1989, **218:**465-474.
265. Usui T, Ogawa H, Petrini JHJ: **A DNA Damage Response Pathway Controlled by Tel1 and the Mre11 Complex.** *Molecular Cell* 2001, **7:**1255-1266.
266. Wyrick JJ, Holstege FC, Jennings EG, Causton H, Shore D, Grunstein M, Lander ES, Young RA: **Chromosomal landscape of nucleosome-dependent gene expression and silencing in yeast.** *Nature* 1999, **402:**418-421.
267. Redon C, Pilch D, Rogakou E, Sedelnikova O, Newrock K, Bonner W: **Histone H2A variants H2AX and H2AZ.** *Current Opinion in Genetics & Development* 2002, **12:**162-169.
268. Ren Q, Gorovsky MA: **Histone H2A.Z Acetylation Modulates an Essential Charge Patch.** *Molecular Cell* 2001, **7:**1329-1335.
269. Madigan JP, Chotkowski HL, Glaser RL: **DNA double-strand break-induced phosphorylation of Drosophila histone variant H2Av helps prevent radiation-induced apoptosis.** *Nucl Acids Res* 2002, **30:**3698-3705.
270. Millar CB, Xu F, Zhang K, Grunstein M: **Acetylation of H2AZ Lys 14 is associated with genome-wide gene activity in yeast.** *Genes Dev* 2006, **20:**711-722.

VITA

Xochitl Chamorro Morgan, the daughter of Michael Jack and Sherry Carr, was born in Grand Rapids, Michigan, USA in 1980. She attended Rockford High School in Rockford, Michigan. In 1996, she enrolled at Simon's Rock College of Bard in Great Barrington, MA; in 1998 she completed both her high school diploma and an Associate of the Arts in liberal arts. She received a Bachelor of Science in Biology from Bowling Green State University in Bowling Green, Ohio in May of 2001. That fall, she started graduate work at the Institute for Cell and Molecular Biology at the University of Texas at Austin. During graduate school she was a co-author on one paper and the first author of two manuscripts, the first of which has been peer-reviewed and published (*BMC Bioinformatics,* 8:445, Nov 2007) and the second of which is currently in preparation. These two works constitute the bulk of this work.

Permanent Address: c/o Jack and Sherry Carr, 1826-11 Mile Rd NE, Comstock Park, MI 49321.

This dissertation was typed by the author.